Dirk Ostwald, Evgeniya Kirilina, Ludger Starke, Felix Blankenburg

A tutorial on Variational Bayes for latent linear stochastic time-series models – Supplement

This document contains the mathematical details that underlie the developments in main tutorial as submitted to the Journal of Mathematical Psychology in its first revision. The current version was finalized August 16th 2013 and will undergo additional proof-reading and revision over the course of autumn 2013. We use the following notation to disentangle mathematical derivations from the main text

○ Start of derivation

□ End of derivation and return to the main text

Outline

1 On Notation

In this section, conventions for the mathematical notation used in the tutorial and supplementary material are discussed. In general, standard mathematical notation is employed, with some concessions to the simplified notations commonly used in statistics and machine learning. The general aim of the notation has been the attempt to balance mathematical rigor and practical applicability of the presented framework.

Sets and Mappings

For analytical discussions sets S that comprise elements sharing a set-specific property p are denoted in the form

$$S = \{s | s\ has\ the\ property\ p\} \tag{1.1}$$

A function (mapping) f is generally specified in the form

$$f: D \to R, x \mapsto f(x) \tag{1.2}$$

where the set D denotes the domain of the function f, the set R denotes the range of the function f, and '$\mapsto$' denotes the mapping of the domain element $x \in D$ onto the range element $f(x) \in R$. In concordance with contemporary mathematical language, the terms 'function' and 'mapping' are used interchangeably in this tutorial.
It is important to note that this (mathematical) notation distinguishes between the function f proper, and elements $f(x)$ of its range. Usually in statements as the above, the specification of the function abbreviation, its domain and its range, is followed by a definition of the functional form linking the domain elements to the range elements (for example for a quadratic function: $f(x) \coloneqq x^2$).

Derivatives and Integrals

The derivative of a function of a single variable is usually denoted in 'operator' form as

$$\frac{d}{dx} f: D \to R, x \mapsto \frac{d}{dx} f(x) \tag{1.3}$$

where

$$\frac{d}{dx} f(x)|_{x=a} \tag{1.4}$$

denotes the value of the derivative in $x = a$. The partial derivative of a function of multiple variables $x_i,\ i = 1, \dots, n$ with respect to variable x_i is denoted as

$$\frac{\partial}{\partial x_i} f: D \to R, x \mapsto \frac{\partial}{\partial x_i} f(x) \tag{1.5}$$

Sometimes the abbreviated bar notation for the derivative, $f'(x)$, is also used.

Integrals of a function f are usually denoted as

$$\int_a^b f(\xi)\, d\xi \tag{1.6}$$

Often, if the domain of integration is left unspecified, or implicitly understood as comprising the entire real line, the integral boundaries a, b are omitted. Likewise, if the variable of integration is n-dimensional, the respective n-time integration is left implicit, e.g. for a function $f: \mathbb{R}^3 \to \mathbb{R}$, the integral on $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] \subset \mathbb{R}^3$

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} f(x_1, x_2, x_3)\, dx_1\, dx_2 dx_3 \tag{1.7}$$

is usually denoted as

$$\int f(x)dx \tag{1.8}$$

Probabilistic concepts

With respect to probabilistic concepts, the applied probabilistic notation prevalent in physics, machine learning, and statistics is employed. Specifically, probability spaces comprising an outcome set Ω, a σ-algbra $\mathcal{A}$ and a probability measure P, are not explicitly defined. Likewise, random variables will usually not be defined as measurable mappings from one measure space to another in the form

$$X\colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}'), \omega \mapsto X(\omega) \tag{1.9}$$

Rather, the exposition focuses on image measures, or the distributions of random variables, written as $P(X) \coloneqq P_X(X)$, and generally defined by

$$P(X) \coloneqq P_X\{X \in A'\} \coloneqq P\{\omega \in \Omega | X(\omega) \in A'\} \ (A \in \mathcal{A}') \tag{1.10}$$

For example, the statement

$$p(x = 1) = 0.1 \tag{1.11}$$

should be read as 'the probability that the random variable x takes on a value of 1 is 0.1'. More generally, arbitrary values that a random value may take on are denoted by a superscripted asterisk, e.g. $p(x = x^*)$. An explicit distinction between probabilistic statements as frequency limits or degrees of beliefs is omitted. In concordance with the notations in machine learning textbooks, the explicit differentiation between probability distributions, probability density functions, and probability mass functions is omitted and non-capital letters are used to denote random variables and probabilities. Given the focus on Gaussian random variables in this tutorial, all probability distributions employed will actually be probability density functions, and the learning of the parameters of probability distributions should be understood as probability density function parameter learning. Finally, while some univariate examples are discussed, the theory is in general developed for random vectors, i.e. vectors of random variables of the form $x = (x_1, \dots, x_d)^T$ where the entries $x_1, \dots, x_d$ represent random variables, and $d \in \mathbb{N}$ the dimensionality of the vector. This treatment subsumes the random variable case with $d = 1$. The most important aspect of this is to keep in mind that the covariances of vectors are not positive scalars as in the random variable case, but symmetric positive-semidefinite matrices. Occasionally, colloquial reference will be made to "random variables" in cases where the random vector concept is equally appropriate.

As a notational example, the following conventions for writing bivariate and marginal probability distributions of random vectors x and y are used:

$$p(x, y)\colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+, \ (x = x^*, y = y^*) \mapsto p(x = x^*, y = y^*) \tag{1.12}$$

for a bivariate probability distribution, where $\mathcal{X}$ and $\mathcal{Y}$ denote the ranges of x and y, respectively, and

$$p(x)\colon \mathcal{X} \longrightarrow \mathbb{R}_+, \ x = x^* \mapsto p(x = x^*) = \int p(x = x^*, y)dy \tag{1.13}$$

for the corresponding marginal distribution over x. With respect to conditional probability distributions, denoted here as,

$$p(x|y)\colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+, \ (x = x^*, y = y^*) \mapsto p(x = x^*|y = y^*) = \frac{p(x=x^*, y=y^*)}{\int p(x, y=y^*)dx} \tag{1.14}$$

an important notational convention is employed to distinguish between probability distributions parameterized by fixed quantities, i.e. 'nonrandom' variables and probability distributions conditioned on other random variables.

Specifically, $p_\theta(x)$ is written for a probability distribution over the random variable x, which is parameterized by the fixed (nonrandom) variable θ. Likewise, $p_\vartheta(x|\theta)$ is written for a probability distribution that is conditioned on the random variable θ and parameterized by the fixed (nonrandom) variable ϑ. Finally, in the case of multivariate probability distributions, conditional distributions over subsets of variables are often referred to as conditional marginal distributions.

Due to the fact that the parameters of a Gaussian distribution are equivalent with its first central moments, namely its expectation and covariance, frequent use is made of expectation and covariance operations on random vectors. Here, the notations $\mathbb{E}_{p(x)}\big(f(x)\big)$ and $\langle f(x)\rangle_{p(x)}$ are used interchangeably to denote the expectation operation of a function f of the random vector $x \in \mathcal{X}$ under the probability distribution p:

$$\langle\cdot\rangle_{\cdot}: \mathcal{F}(x) \times \mathcal{P}(x) \to \mathbb{R},\ (p(x), f(x)) \mapsto \langle f(x)\rangle_{p(x)} \coloneqq \int f(x)p(x)dx \tag{1.15}$$

and

$$\mathbb{E}_{\cdot}(\cdot): \mathcal{F}(x) \times \mathcal{P}(x) \to \mathbb{R},\ (p(x), f(x)) \mapsto \mathbb{E}_{p(x)}\big(f(x)\big) \coloneqq \int f(x)p(x)dx \tag{1.16}$$

where $\mathcal{F}(x)$ and $\mathcal{P}(x)$ denote unspecified[1] function and probability distribution sets over x and the integration is performed over the range $\mathcal{X}$ of x. The reason for the use of these interchangeable notations is that (in the opinion of the authors) the bracketed notation $\langle f(x)\rangle_{p(x)}$ emphasizes the 'operator' characteristic of an expectation slightly more, and is hence used mainly in algebraic manipulations, while the notation $\mathbb{E}_{p(x)}\big(f(x)\big)$ stresses the 'single value' characteristic of an expectation or conditional expectation slightly more and is hence used predominantly in the statement of results rather than calculations. Covariance operations are usually denoted by the symbol $\mathbb{C}$ in the form

$$\mathbb{C}(\cdot,\cdot): \mathcal{P}(x,y) \to \mathbb{R}_+, p(x,y) \mapsto \mathbb{C}_{p(x,y)}(x,y) \coloneqq\ \mathbb{E}_{p(x,y)}\left(\Big(x - \mathbb{E}_{p(x)}(x)\Big)\Big(y - \mathbb{E}_{p(y)}(y)\Big)^T\right) \tag{1.17}$$

which for a random vectors of dimensionality $d \in \mathbb{N}$ refers to a symmetric, positive-semidefinite matrix of dimensionality $d \times d$. The probability distribution subscripts in the abbreviations $\langle x\rangle_{p(x)}, \mathbb{E}_{p(x)}(x)$ and $\mathbb{C}_{p(x,y)}(x,y)$ are also often omitted, if the respective distributions is clear from the context.

For random vectors x, y, i.e. vectors of random variables $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d), d \in \mathbb{N}$ and scalars $a, b, c, d \in \mathbb{R}$, some standard rules for manipulations involving expectations, variances and covariances are listed below.

Expectation of an expectation of a RV

$$\mathbb{E}\big(\mathbb{E}(x)\big) = \mathbb{E}(x) \tag{1.18}$$

Expectation of the product of a scalar and an RV

$$\mathbb{E}(ax) = a\mathbb{E}(x) \tag{1.19}$$

Expectation of the sum of two RVs

$$\mathbb{E}(x + y) = \mathbb{E}(x) + \mathbb{E}(y) \tag{1.20}$$

Alternative expression for the variance of an RV

[1] The spaces $\mathcal{F}$ and $\mathcal{P}$ are left unspecified to eschew a formal discussion of measure theoretic concepts. The same reasoning is behind the simplified probability concept notations discussed.

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \tag{1.21}$$

Alternative expression for the covariance of RVs x, y

$$\mathbb{C}(x, y) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y) \tag{1.21}$$

Covariance of an RV x with itself

$$\mathbb{C}(x, x) = \mathbb{V}(x) \tag{1.22}$$

Variance of a RV under scalar multiplication and addition

$$\mathbb{V}(ax + c) = a^2\mathbb{V}(x) \tag{1.23}$$

Note that (1.18) – (1.23) are readily transferred to the random vector case, if the corresponding rules of vector calculus are respected.

Finally, the letters μ, Σ and σ are usually used to denote the expectation and (co)variance of Gaussian random variables and should not be confused with the expectation and (co)variance operations introduced above. Gaussian distributions are denoted by $N(x; \mu, \Sigma)$, where x indicates the Gaussian random vector, and μ, Σ its parameters (see Supplement B for details).

Statistical concepts

With respect to the notation of statistical concepts, we note that we will mainly be dealing with iterative estimation schemes. We will denote the value that an estimated parameter takes on the ith iteration of some iterative parameter estimation scheme is denoted by a superscript (i), such as $\theta^{(i)}, \mu^{(i)}, A^{(i)}, \Sigma_x^{(i)}$. The superscript notation thus makes the usual 'hat' notation for parameter estimates ($\hat{\theta}$, for example) redundant, which is hence omitted.

Table 1 summarizes the notational conventions used in the tutorial and supplementary material.

Notation	Meaning
$f: D \to R, x \mapsto f(x)$	f denotes a mapping of elements x in its domain D onto elements $f(x)$ of its range R.
$\frac{d}{dx} f: D \to R, x \mapsto \frac{d}{dx} f(x)$	$\frac{d}{dx} f$ denotes the derivative of f, which corresponds to a mapping of x onto $\frac{d}{dx} f(x)$.
$\frac{\partial}{\partial x_i} f: D \to R, x \mapsto \frac{\partial}{\partial x_i} f(x)$	For functions of multiple variables $x_1, \dots, x_n$ (e.g., $D \subset \mathbb{R}^n$), $\frac{\partial}{\partial x_i} f$ denotes the partial derivative of f with repsect to x_i.
$\int_a^b f(\xi)\, d\xi$	Integral notation a, b indicate the lower and upper boundaries and ξ the integration variable.
$p(x, y): \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ $(x = x^*, y = y^*) \mapsto p(x = x^*, y = y^*)$	Joint probability distribution notation as prevalent in applied mathematical texts. In the context of this tutorial, $p(x, y)$ usually refers to a probability density function.
$p(x): \mathcal{X} \longrightarrow \mathbb{R}_+ x = x^* \mapsto$ $p(x = x^*) \coloneqq \int p(x = x^*, y) dy$	Marginal probability distribution as obtained by integration ("marginalization") of continuous probability distributions.
$p(x\|y): \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+, (x = x^*, y = y^*) \mapsto$ $p(x = x^* \| y = y^*) \coloneqq \frac{p(x=x^*, y=y^*)}{\int p(x, y=y^*) dx}$	Definition of conditional probability distibutions. The conditional distribution of x given $y = y^*$ equals the normalized joint probability distribution of x and $y = y^*$.
$\langle \cdot \rangle_\cdot: \mathcal{F}(x) \times \mathcal{P}(x) \to \mathbb{R}$ $(p(x), f(x)) \mapsto \langle f(x) \rangle_{p(x)} \coloneqq \int f(x) p(x) dx$	Notation of the expectation operation for functions f of random variables x.
$\mathbb{E}_\cdot(\cdot): \mathcal{F}(x) \times \mathcal{P}(x) \to \mathbb{R}$ $(p(x), f(x)) \mapsto \mathbb{E}_{p(x)}\big(f(x)\big) \coloneqq \int f(x) p(x) dx$	Alternative notation of the expectation for functions f of random variables x.
$\mathbb{C}(\cdot, \cdot): \mathcal{P}(x, y) \to \mathbb{R}_+, p(x, y) \mapsto$ $\mathbb{C}_{p(x,y)}(x, y) \coloneqq \mathbb{E}_{p(x,y)}\left(\big(x - \mathbb{E}_{p(x)}(x)\big)\big(y - \mathbb{E}_{p(y)}(y)\big)^T\right)$	Notation of the covariance for random variables x.
μ, Σ, σ	Symbols for the parameters of Gaussian distributions.
$\mu_{x\|y}, \Sigma_{x\|y}, \sigma_{x\|y}$	Notation for "conditional parameters." The parameter of interest is subscripted with the random variable whose distribution it governs (x) and which is conditioned on another random variable (y).
$\theta^{(i)}, \mu^{(i)}, A^{(i)}, \Sigma_x^{(i)}$	Symbols for the value of parameter estimates on the ith iteration of an iterative parameter estimation algorithm.
$\Sigma \succ 0$	Σ is a positive-definite matrix
$\mathcal{C}^k(\mathbb{R})$	The space of k-times continuously differentiable functions on $\mathbb{R}$
$\|\Sigma\|$	Determinant of the matrix Σ
$x \sim p(x)$	x is distributed according to $p(x)$

Table 1 Notational Conventions of the Tutorial and Supplement

2 Some Properties of Gaussian and Gamma Distributions

The analytical properties of multivariate Gaussian distributions are of central importance for the VB approach for LGSSMs In this section, we summarize some of these properties in theorem form. Additionally, an overview is provided in Table 2 and Figure 1.

Definition and central moments of the Gaussian

A Gaussian distribution over a random vector, i.e., a vector of random variables, $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ is defined by the probability density function

$$p_{\mu,\Sigma}(x) \coloneqq N(x;\mu,\Sigma) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{2.1}$$

Here, $\mu \in \mathbb{R}^d$ and the positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succ 0$ are the fixed parameters of the distribution as indicated by the subscript and semicolon notation in $p_{\mu,\Sigma}(x)$ and $N(x;\mu,\Sigma)$, respectively. If μ and Σ are themselves random variables, the statement above is written $p(x|\mu,\Sigma) = N(x|\mu,\Sigma)$ and takes on its parameterized form only for concrete values of the random variables $\mu = \mu^*$ and $\Sigma = \Sigma^*$, i.e. $p(x|\mu = \mu^*, \Sigma = \Sigma^*) = N(x;\mu^*,\Sigma^*)$. For the case of a univariate random variable $x \in \mathbb{R}$, the probability density function is usually expressed as

$$p_{\mu,\sigma^2}(x) \coloneqq N(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{2.2}$$

with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$. The first central moment or the expectation of normal distribution is given by

$$\mathbb{E}_{p_{\mu,\Sigma}(x)}(x) = \langle x\rangle_{p_{\mu,\Sigma}(x)} = \mu \tag{2.3}$$

i.e., the parameter $\mu \in \mathbb{R}^d$ of a multivariate normal distribution $N(x;\mu,\Sigma)$ is equivalent to the expectation of the identity function of x under the probability distribution $N(x;\mu,\Sigma)$. The second central moment or the covariance of a normal distribution is given by

$$\mathbb{C}_{p_{\mu,\Sigma}(x)}(x,x) = \mathbb{E}_{p_{\mu,\Sigma}(x)}\left(\left(x - \mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\right)\left(x - \mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\right)^T\right) = \Sigma \tag{2.4}$$

i.e., the parameter $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succ 0$ of a multivariate normal distribution $N(x;\mu,\Sigma)$ is equivalent to the covariance of the identity function of x under the probability distribution $N(x;\mu,\Sigma)$. These two equivalences allow the expectation of the square of x under $p_{\mu,\Sigma}(x)$ to be expressed in terms of the expectation and covariance of x under $p_{\mu,\Sigma}(x)$, an important property, which is exploited on several ocassions in later sections. Specifically, because of

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \tag{2.5}$$

one has for normally distributed random vectors $x \in \mathbb{R}^d$:

$$\begin{aligned}\mathbb{E}_{p_{\mu,\Sigma}(x)}(xx^T) &= \mathbb{V}_{p_{\mu,\Sigma}(x)}(x) + \mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\mathbb{E}_{p_{\mu,\Sigma}(x)}(x)^T \\ &= \mathbb{C}_{p_{\mu,\Sigma}(x)}(x,x) + \mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\mathbb{E}_{p_{\mu,\Sigma}(x)}(x)^T \\ &= \langle xx^T\rangle_{p_{\mu,\Sigma}(x)} + \langle x\rangle_{p_{\mu,\Sigma}(x)}\langle x\rangle^T_{p_{\mu,\Sigma}(x)} \\ &= \Sigma + \mu\mu^T\end{aligned} \tag{2.6}$$

The Completing-the-square theorem for the Gaussian distribution

The completing-the-square theorem states that if $x \in \mathbb{R}^d, b \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}, A \succ 0$, then

$$\exp\left(-\frac{1}{2}x^T Ax + b^T x\right) = N(x; A^{-1}b, A^{-1})\sqrt{|(2\pi A)^{-1}\ |}\exp\left(\frac{1}{2}b^T A^{-1}b\right) \tag{2.7}$$

The completing-the-square theorem allows identifying the parameters of a Gaussian probability density function of $x \in \mathbb{R}^d$ based on a quadratic form $x \in \mathbb{R}^d$, given by the argument of the exponent on the left hand side of the theorem.

○ It is verified by substitution of $N(x; A^{-1}b, A^{-1})$ on the the right-hand side of the identity above as follows:

$$N(x; A^{-1}b, A^{-1})\sqrt{|(2\pi A)^{-1}\ |}\exp\left(\frac{1}{2}b^T A^{-1}b\right) \tag{2.8}$$

$$= \frac{\sqrt{|(2\pi A)^{-1}\ |}}{\sqrt{|(2\pi A)^{-1}\ |}}\exp\left(-\frac{1}{2}(x - A^{-1}b)^T A(x - A^{-1}b) + \frac{1}{2}b^T A^{-1}b\right)$$

$$= \exp\left(-\frac{1}{2}\left(x^T Ax - x^T AA^{-1}b - b^T A^{-1^T} Axb^T A^{-1^T} AA^{-1}b\right) + \frac{1}{2}b^T A^{-1}b\right)$$

$$= \exp\left(-\frac{1}{2}(x^T Ax - x^T b - b^T A^{-1}Ax + b^T A^{-1}b) + \frac{1}{2}b^T A^{-1}b\right)$$

$$= \exp\left(-\frac{1}{2}x^T Ax + \frac{1}{2}b^T x + \frac{1}{2}b^T x - \frac{1}{2}b^T A^{-1}b + \frac{1}{2}b^T A^{-1}b\right)$$

$$= \exp\left(-\frac{1}{2}x^T Ax + b^T x\right)$$

□

The Linear transformation theorem for the Gaussian distribution

The "linear transformation theorem" states that if $x \sim N(x; \mu_x, \Sigma_x)$, where $\mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d\times d}, \Sigma_x \succ 0$ $\varepsilon \sim N(\varepsilon; \mu_\varepsilon, \Sigma_\varepsilon)$, where $\varepsilon, \mu_\varepsilon \in \mathbb{R}^d, \Sigma_\varepsilon \succ 0$, $\mathbb{C}(x, \varepsilon) = \big(\mathbb{C}(\varepsilon, x)\big)^T = 0 \in \mathbb{R}^{d\times d}$ and $A \in \mathbb{R}^{d\times d}$ is a matrix, then

$$Ax + \varepsilon =: y \sim N\big(y; \mu_y, \Sigma_y\big) \tag{2.9}$$

where $y \in \mathbb{R}^d$, $\mu_y \in \mathbb{R}^d$, $\Sigma_y \in \mathbb{R}^{d\times d}$, Σ_y and specifically,

$$\mu_y = A\mu_x + \mu_\varepsilon \text{ and } \Sigma_y = A\Sigma_x A^T + \Sigma_\varepsilon \tag{2.10}$$

The proof that y is indeed a Gaussian random vector capitalizes on the transformation theorem for probability density functions and is suppressed here for brevity.

○ If it is acknowledged that y is a Gaussian random vector, the expressions for the parameters of this Gaussian can be derived in an instructive manner from the fact that the parameters of a Gaussian correspond to its first two central moments and the general rules for manipulating expectancies and covariances as follows:

$$\mu_y \coloneqq \mathbb{E}(y) = \mathbb{E}(Ax + \varepsilon) = \mathbb{E}(Ax) + \mathbb{E}(\varepsilon) = A\mathbb{E}(x) + \mathbb{E}(\varepsilon) = A\mu_x + \mu_\varepsilon \tag{2.11}$$

and

$$\Sigma_y =: \mathbb{E}\left(\big(y - \mathbb{E}(y)\big)\big(y - \mathbb{E}(y)\big)^T\right) \tag{2.12}$$

$$= \mathbb{E}\left(\left(Ax + \varepsilon - \mathbb{E}(Ax + \varepsilon)\right)\left(Ax + \varepsilon - \mathbb{E}(Ax + \varepsilon)\right)^T\right)$$

$$= \mathbb{E}\left(\left(Ax + \varepsilon - A\mathbb{E}(x) - \mathbb{E}(\varepsilon)\right)\left(Ax + \varepsilon - A\mathbb{E}(x) - \mathbb{E}(\varepsilon)\right)^T\right)$$

$$= \mathbb{E}\left(\left(A\left(x - \mathbb{E}(x)\right) + \left(\varepsilon - \mathbb{E}(\varepsilon)\right)\right)\left(A\left(x - \mathbb{E}(x)\right) + \left(\varepsilon - \mathbb{E}(\varepsilon)\right)\right)^T\right)$$

$$= \mathbb{E}\left(\left(A\left(x - \mathbb{E}(x)\right) + \left(\varepsilon - \mathbb{E}(\varepsilon)\right)\right)\left(\left(x - \mathbb{E}(x)\right)^T A^T + \left(\varepsilon - \mathbb{E}(\varepsilon)\right)^T\right)\right)$$

$$= A\mathbb{E}\left(\left(x - \mathbb{E}(x)\right)\left(x - \mathbb{E}(x)\right)^T\right)A^T + A\mathbb{E}\left(\left(x - \mathbb{E}(x)\right)\left(\varepsilon - \mathbb{E}(\varepsilon)\right)^T\right)$$

$$+\mathbb{E}\left(\left(\varepsilon - \mathbb{E}(\varepsilon)\right)\left(x - \mathbb{E}(x)\right)^T\right)A^T + \mathbb{E}\left(\left(\varepsilon - \mathbb{E}(\varepsilon)\right)\left(\varepsilon - \mathbb{E}(\varepsilon)\right)^T\right)$$

$$= A\mathbb{C}(x, x)A^T + A\mathbb{C}(x, \varepsilon) + \mathbb{C}(\varepsilon, x)A^T + \mathbb{C}(\varepsilon, \varepsilon)$$

$$= A\Sigma_x A^T + A \cdot 0 + 0 \cdot A^T + \Sigma_\varepsilon$$

$$= A\Sigma_x A^T + \Sigma_\varepsilon$$

where in the penultimate line the independence property $\mathbb{C}(x, \varepsilon) \coloneqq 0$ of x and ε was exploited.

□

The Marginalization and conditioning theorem for the Gaussian distribution

The marginalization and conditioning theorem states that, if $z \sim N(z; \mu, \Sigma), z, \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}, \Sigma \succ 0$ and

$$z \coloneqq \begin{pmatrix} x \\ y \end{pmatrix}, x \in \mathbb{R}^m, y \in \mathbb{R}^{d-m} \tag{2.13}$$

with corresponding partitions of the parameters and precision matrix $\Lambda = \Sigma^{-1} \in \mathbb{R}^{d\times d}$, i.e., with

$$\mu \coloneqq \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix} \tag{2.14}$$

then

$$p(x) = N(x; \mu_x, \Sigma_{xx}) \tag{2.15}$$

Further,

$$p(x|y) = N\left(x; \mu_{x|y}, \Sigma_{x|y}\right) \tag{2.16}$$

where

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \text{ and } \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \tag{2.17}$$

- *Preliminaries*

To verify the marginalization and conditioning theorem, we follow (Bishop, 2007). The expressions for the conditional mean and conditional covariance can be verified using two intermediate steps. We first obtain some helpful relations between the elements of matrices Σ and Λ. Using the fact that $\Lambda = \Sigma^{-1}$ we have:

$$\begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{pmatrix}\begin{pmatrix}\Lambda_{xx} & \Lambda_{xy}\\ \Lambda_{yx} & \Lambda_{yy}\end{pmatrix} = I \tag{2.18}$$

Based on (2.18) one can infer the following identities

$$\Lambda_{xx} - \Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx} = \Sigma_{xx}^{-1} \tag{2.19}$$

$$\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} = \Lambda_{xx}^{-1} \tag{2.20}$$

and

$$\Lambda_{yy}^{-1}\Lambda_{yx} = -\Sigma_{yx}\Sigma_{xx}^{-1} \tag{2.21}$$

We next consider the joint probability distribution

$$p(x,y) = p(z) = N(z;\mu,\Sigma) \tag{2.22}$$

and define

$$a \coloneqq (x-\mu_x) \text{ and } b \coloneqq (y-\mu_y) \tag{2.23}$$

Then the quadratic form in the exponent of $p(x,y)$ can be rewritten as:

$$-\tfrac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu) = -\tfrac{1}{2}\begin{pmatrix}a & b\end{pmatrix}\begin{pmatrix}\Lambda_{xx} & \Lambda_{xy}\\ \Lambda_{yx} & \Lambda_{yy}\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} \tag{2.24}$$

$$= -\tfrac{1}{2}\left(a^T\Lambda_{xx}a + a^T\Lambda_{xy}b + b^T\Lambda_{yx}a + b^T\Lambda_{yy}b\right)$$

$$= -\tfrac{1}{2}\left(a^T\Lambda_{xx}a + a^T\Lambda_{xy}\Lambda_{yy}\Lambda_{yy}^{-1}b + b^T\Lambda_{yy}\Lambda_{yy}^{-1}\Lambda_{yx}a + b^T\Lambda_{yy}b\right)$$

$$= -\tfrac{1}{2}a^T\left(\Lambda_{xx} - \Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx}\right)a - \tfrac{1}{2}\left(b + \Lambda_{yy}^{-1}\Lambda_{yx}a\right)^T\Lambda_{yy}\left(b + \Lambda_{yy}^{-1}\Lambda_{yx}a\right)$$

Further, using the relations (2.18) – (2.20), (2.24) can be rewritten as either

$$-\tfrac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu) = -\tfrac{1}{2}(x-\mu_x)^T\Sigma_{xx}^{-1}(x-\mu_x) - \tfrac{1}{2}\left(y-\mu_y-\Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x)\right)^T\Lambda_{yy}\left(y-\mu_y-\Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x)\right) \tag{2.25}$$

or

$$-\tfrac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu) = -\tfrac{1}{2}\left(x-\mu_x-\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)^T\Lambda_{xx}\left(x-\mu_x-\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right) - \tfrac{1}{2}(y-\mu_y)^T\Sigma_{yy}^{-1}(y-\mu_y) \tag{2.26}$$

□

- *Verification of (2.15)*

To obtain the marginal distribution $p(x)$, the joint distribution $p(x,y)$ has to be integrated over y. To this end, we make use of (2.25)

$$p(x) = \int p(x,y)dy \tag{2.27}$$

$$= (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(x-\mu_x)^T\Sigma_{xx}^{-1}(x-\mu_x)\right)$$

$$\cdot \int \exp\left(-\frac{1}{2}\left(y-\mu_y-\Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x)\right)^T \Lambda_{yy}\left(y-\mu_y-\Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x)\right)\right)dy$$

$$= (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(x-\mu_x)^T\Sigma_{xx}^{-1}(x-\mu_x)}(2\pi)^{\frac{m}{2}}\left|\Lambda_{yy}\right|^{-\frac{1}{2}}$$

$$= N(x;\mu_x,\Sigma_{xx})$$

This provides verification of the statement (2.15).

□

- *Verification of (2.16)*

We now consider the joint probability distribution $p(x, y = y^*)$, which, upon normalization, corresponds to the conditional distribution $p(x|y = y^*)$ via

$$p(x|y=y^*) \coloneqq \frac{p(x,y=y^*)}{p(y=y^*)} \tag{2.28}$$

Rewriting the unnormalized conditional distribution using quadratic form (2.26) and summarizing all terms independent of x independent terms in a constant C, one obtains

$$p(x|y=y^*) = C\cdot\exp\left(-\frac{1}{2}\left(x-\mu_x-\Sigma_{xy}\Sigma_{yy}^{-1}(y^*-\mu_y)\right)^T\Lambda_{xx}\left(x-\mu_x-\Sigma_{xy}\Sigma_{yy}^{-1}(y^*-\mu_y)\right)\right) \tag{2.29}$$

Implicitly assuming appropriate normalization we see that conditional distribution has a Gaussian form with parameters

$$\Sigma_{x|y^*} \coloneqq \Lambda_{xx}^{-1} = \Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \tag{2.30}$$

and

$$\mu_{x|y^*} \coloneqq \mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y^*-\mu_y) \tag{2.31}$$

□

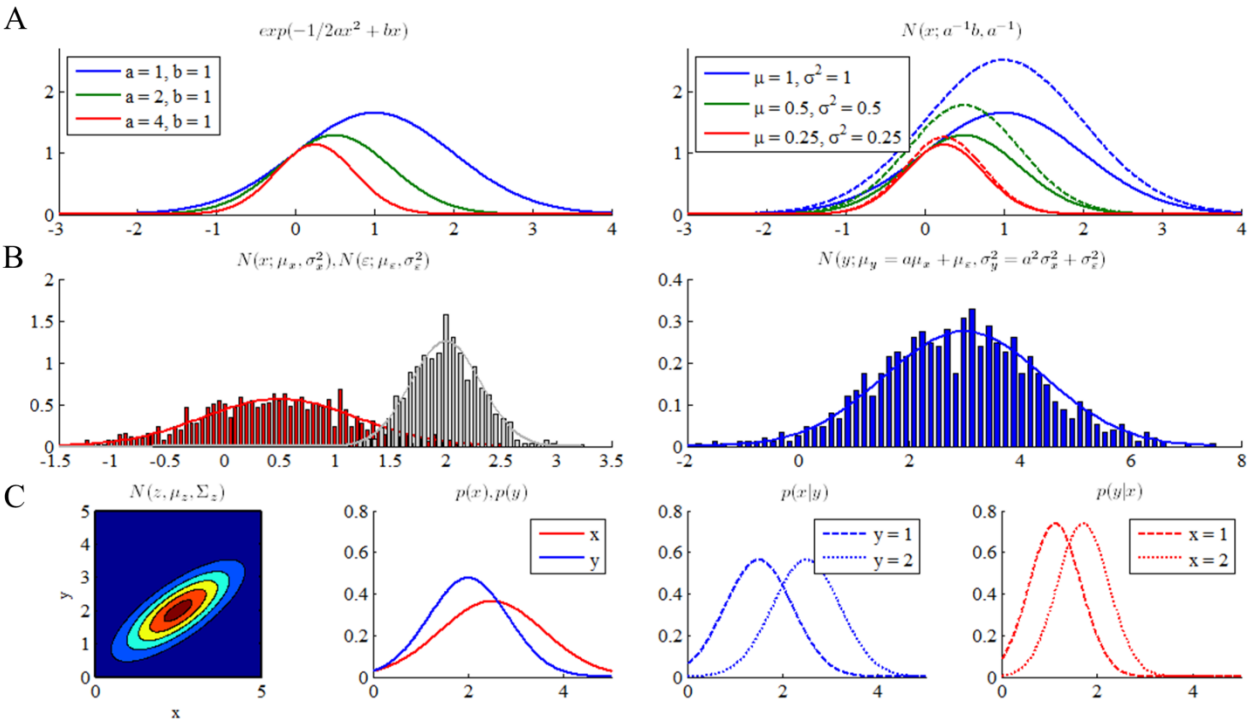

Figure 1. Visualization of properties of the Gaussian distribution. Figure 2A visualizes the completing-the-square theorem. The left-hand panel depicts three different exponential quadratic forms in x with parameters a and b as noted in the legend. For the same parameter settings, the right-hand panel depicts the evaluation of the first multiplicative term of the right-hand side of the completing-the-square theorem, i.e., the Gaussian form with parameters $\mu = a^{-1}b$ and $\sigma^2 = a^{-1}$ as dashed curves. The continuous curves represent these forms upon multiplication with the normalization factor, rendering the ensuing curves equivalent to the exponential quadratic forms in x depicted in the left-hand panel. Figure 2B visualizes the linear transformation theorem both on the analytical and realized level. The left-hand panel depicts the functional forms of two Gaussian distributions over variables x and ε as red and grey continuous curves, respectively. Additionally, a histogram estimate of both density functions based on $N = 1000$ samples is depicted as a bar graph. The right-hand panel depicts the ensuing functional form of the distribution of the random variable $x + \varepsilon$ as well as the result of a histogram estimate of the addition of the samples obtained for the left-hand panel. Figure 2C visualizes the Gaussian marginalization and conditioning theorem for a bivariate Gaussian distribution over $z = (x, y)^T$. The leftmost panel depicts the functional form of the joint distribution over x and y. The marginal distributions resulting from the application of the marginalization equations are depicted in the second panel from the left. For values $y = 1$ and $y = 2$, the third panel from the left depicts the conditional marginal distributions $p(x|y)$ as dashed and dotted lines, respectively; likewise, for values $x = 1$ and $x = 2$, the rightmost panel depicts the conditional marginal distributions $p(y|x)$ as dashed and dotted lines, respectively.

Definition and notation of the multivariate Gaussian distribution for $x \in \mathbb{R}^d$

$$p_{\mu,\Sigma}(x) \coloneqq N(x;\mu,\Sigma) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Definition and notation of the univariate Gaussian distribution $x \in \mathbb{R}$

$$p_{\mu,\sigma^2}(x) \coloneqq N(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Expectation and covariance equalities and notation for multivariate Gaussian distributions

$$\mathbb{E}_{p_{\mu,\Sigma}(x)}(x) = \langle x\rangle_{p_{\mu,\Sigma}(x)} = \mu$$

$$\mathbb{C}_{p_{\mu,\Sigma}(x)}(x,x) = \mathbb{E}_{p_{\mu,\Sigma}(x)}\left(\left(x-\mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\right)\left(x-\mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\right)^T\right) = \Sigma$$

General variance property and application to the multivariate Gaussian distribution

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \Rightarrow \mathbb{E}_{p_{\mu,\Sigma}(x)}(xx^T) = \mathbb{C}_{p_{\mu,\Sigma}(x)}(x,x) + \mathbb{E}_{p_{\mu,\Sigma}(x)}(x)\mathbb{E}_{p_{\mu,\Sigma}(x)}(x)^T = \Sigma + \mu\mu^T$$

"Completing-The-Square Theorem"

The completing-the-square theorem states that if $x \in \mathbb{R}^d, b \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}, A \succ 0$ then

$$\exp\left(-\frac{1}{2}x^TAx - b^Tx\right) = N(x;A^{-1}b,A^{-1})\sqrt{|(2\pi A)^{-1}|}\exp\left(\frac{1}{2}b^TA^{-1}b\right)$$

"Linear Transformation Theorem"

The "linear transformation theorem" states that if $x \sim N(x;\mu_x,\Sigma_x)$, where $,\mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d\times d}, \Sigma_x \succ 0$ $\varepsilon \sim N(\varepsilon;\mu_\varepsilon,\Sigma_\varepsilon)$,

where $\varepsilon, \mu_\varepsilon \in \mathbb{R}^d, \Sigma_\varepsilon \succ 0$, $\mathbb{C}(x,\varepsilon) = \left(\mathbb{C}(\varepsilon,x)\right)^T = 0 \in \mathbb{R}^{d\times d}$ and $A \in \mathbb{R}^{d\times d}$ is a matrix then

$$Ax + \varepsilon =: y \sim N(y;\mu_y,\Sigma_y)$$

where $y \in \mathbb{R}^d$, $\mu_y \in \mathbb{R}^d$, $\Sigma_y \in \mathbb{R}^{d\times d}$, Σ_y and specifically

$$\mu_y = A\mu_x + \mu_\varepsilon \text{ and } \Sigma_y = A\Sigma_xA^T + \Sigma_\varepsilon$$

"Marginalization and Conditioning Theorem"

The marginalization and conditioning theorem states that if $z \sim N(z;\mu,\Sigma), z,\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}, \Sigma \succ 0$ and

$$z \coloneqq \begin{pmatrix}x\\y\end{pmatrix}, x \in \mathbb{R}^m, y \in \mathbb{R}^{d-m}$$

with corresponding partitions of the parameters and variance matrix $\Sigma \in \mathbb{R}^{d\times d}$, i.e., with

$$\mu \coloneqq \begin{pmatrix}\mu_x\\\mu_y\end{pmatrix} \text{ and } \Sigma = \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\\Sigma_{yx} & \Sigma_{yy}\end{pmatrix}$$

then

$$p(x) = N(x;\mu_x,\Sigma_{xx}) \text{ and } p(y) = N(y;\mu_y,\Sigma_{yy})$$

Further,

$p(x|y) = N(x;\mu_{x|y},\Sigma_{x|y})$, where $\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)$ and $\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}$

Table 2 Some Properties of the Gaussian distribution

The Gamma distribution

The Gamma density is considered in its "shape and scale" parameterization given as

$$G(\lambda; a, b) \coloneqq \frac{1}{\Gamma(a)} \frac{1}{b^{\,a}} \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right) \tag{2.32}$$

where a_λ is referred to as the "shape parameter" and b_λ is referred to as the "scale parameter". The expectation and variance of λ under $G(\lambda; a_\lambda, b_\lambda)$ are expressed in terms of the parameters as

$$\mathbb{E}_{G(\lambda;a,b)}(\lambda) = ab \text{ and } \mathbb{V}_{G(\lambda;a,b)}(\lambda) = ab^2 \tag{2.33}$$

The expectation of the logarithm of λ under $G(\lambda; a, b)$ is given by

$$\mathbb{E}_{G(\lambda;a,b)}(\ln \lambda) = \psi(a) + \ln b \tag{2.34}$$

where ψ denotes the digamma function. The KL divergence between two gamma densities $G^q \coloneqq G(\lambda; a^q, b^q)$ and $G^p \coloneqq (\lambda; a^p, b^p)$ is given by (Penny, 2001)

$$\mathcal{KL}(G^q||G^p) = (a^q - 1)\psi(a^q) - \ln b^q - a^q - \ln \Gamma(a^q) + \ln \Gamma(a^p) + a^p \ln b^p - (a^p - 1)(\psi(a^q) + \ln b^q) + \frac{a^q b^q}{b^p} \tag{2.35}$$

3 Wiener Processes, Stochastic Integration, and Euler Maruyama Discretization

The aim of this section is to review a minimum set of concepts from stochastic analysis which forms the background for the embedding of (discrete-time) LGSSMs in the context of (continuous-time) stochastic differential equations as discussed in Sections 2.1 and 5 of the main tutorial. Supplement Section 3 is not meant to provide a comprehensive discussion of stochastic analysis, but rather to collect some pointers to aspects of stochastic analysis the interested reader might pursue. Additionally, we hope to provide some interpretation for equations (1) –(6) of Section 2.1, as well as equations (36) – (37) of the tutorial example in Section 5.

Wiener processes

In this section, the intuition and mathematical formulation of Wiener processes is briefly sketched. For a comprehensive review of stochastic processes in general and Wiener processes in particular, the reader is referred to (Bass, 2011). In the context of the tutorial, Wiener processes form the mathematical model of random fluctuations and the latent level (ref. equation 1 of the main tutorial), which, by means of Euler-Maruyama approximation, results in the Gaussian state space evolution of the LGSSM (ref. equation 2 of the main tutorial). Here, we proceed by introducing the definition of general stochastic processes, the physical intuition of Wiener processes, the definition of Wiener processes as stochastic processes, and finally some properties of Wiener processes required for the discussion of stochastic integration.

General Stochastic Processes

Informally, a general stochastic process can be regarded as a set of random variables $(X_t)_{t\in T}$ indexed by an uncountable index set $T \subseteq \mathbb{R}_+$ that models continuous time. Formally, a general stochastic process is defined as follows: Let $(\Omega, \mathcal{A}, P)$ be a probability space and T an arbitrary nonempty set. Let $X_t\colon (\Omega, \mathcal{A}) \to (\mathcal{X}_t, \mathcal{B}_t)$ be a ($\mathcal{A}$-$\mathcal{B}_t$)- random variable for every $t \in T$. Then

$$X_T \coloneqq (\Omega, \mathcal{A}, P, (X_t)_{t\in T}) \tag{3.1}$$

denotes a general stochastic process with parameter set T.

The uncountability of the parameter set T entails some mathematical difficulties for the definition of the outcome set Ω of a stochastic process, its corresponding σ-field $\mathcal{A}$, and the existence of the probability measure P. These difficulties can be resolved using Kolmogorov's theorem (Bass, 2011), which allows for proofing the existence of a stochastic process based on a consistent set of finite marginal distributions. In other words, the definition of the Wiener processes based on normally distributed increments given below can be related back to the general definition of stochastic processes by means of Kolmogorov's theorem. Below, we often use the simplified notation $X(t)$ for a general stochastic process.

Physical Intuition of Wiener Processes

It is helpful to consider the physical interpretation of a Wiener process to motivate the formal definition below. In its physical interpretation, a realization of one-dimensional Wiener process can be regarded as the projection of the displacement vector of a three-dimensional particle in a fluid onto a selected coordinate axis of $\mathbb{R}^3$ over time. Specifically, $X_0 = 0$ would denote the particle centered on its starting point and X_t the projection of its displacement from the starting point at time t onto e.g., the x-axis. Importantly, the particle is considered large with respect to the molecules constituting the fluid. In principle, based on a set of initial conditions, the movement of all molecules in the fluid is determined by Hamiltonian (i.e., generalized Newtonian) dynamics. However, because the number of molecules is conceived as very large, a statistical or probabilistic view is appropriate. The molecules of the fluid are imagined to collide with the large particle at a time scale that is short with respect to the time scale that the particle is observed on. In this sense, the movement of the large particle in a time interval $[t_1, t_2]$ can be imagined as

being the result of the summation of a large number of independent fluid-molecule-particle collisions, each evoking a minute movement of the particle. Informally, the central limit theorem of probability theorem states that the sum (or average) of a high number of independently distributed random variables (here the minute motions of the particle evoked by each molecule-particle collision) is normally distributed. Thus, displacement in X, i.e. the difference $X_{t_2} - X_{t_1}$ can be considered to be normally distributed, and, because the molecule impacts are conceived as independent and evoking motion in all spatial directions, have an expectation of zero, i.e. $E\left(X_{t_2} - X_{t_1}\right) = 0$. In other words, most likely, due to the complementary nature of the summed impacts between $t_1$and t_2, the particle stays were it is. Further, if the characteristics of the fluid do not change over time (e.g. the fluid is not heated up, causing stronger impacts of the molecules), the distribution of the spatial distance in a time interval $[t_1, t_2]$ should be the same as the distribution of the spatial distance in a time interval $[t_1 + h, t_2 + h]$, with $h > 0$. Finally, the particle motions evoked by the impact of fluid molecules are considered to be independent over disjoint time intervals.

Definition of Wiener Processes[2]

The intuition of a Wiener process realization as a large particle colliding with many small fluid molecules at a short time scale with respect to the observation time scale of the movement of the large particle yields the following set of formal requirements for a mathematical model $(\Omega, \mathcal{A}, P, (X_t)_{t\in T})$ of Wiener processes: $(\Omega, \mathcal{A}, P, (X_t)_{t\in T})$ is a stochastic process with parameter set $T = \mathbb{R}_+^1$, for which the following requirements hold:

(1) *Initial value* $X_0 = 0$ P-almost everywhere (i.e., if there exists $N \subset \Omega$ for which $X_0(\omega_N) \neq 0, \omega_N \in N$, then $P(N) = 0$. In other words, $P\{X_0 = 0\} = P(\{\omega \in \Omega | X_0(\omega) = 1\}) = 1$).
(2) *Normality* For every $t > 0, X_t$ is normally distributed with expectation 0 and variance parameter $v(t)$. Here $v(\cdot)$ denotes the variance of the normal distribution of X_t as a function of t.
(3) *Stationarity and Independence* $(\Omega, \mathcal{A}, P, (X_t)_{t\in T})$ is a stochastic process with stationary increments, i.e. the distribution $P_{X_t - X_s}$ of the increments $X_t - X_s$ $(s, t \in T, s < t)$ only depends on the difference $t - s$, but not on t or s itself. In other words, for $h > 0$, we have $P_{X_{t+h} - X_{s+h}} = P_{X_t - X_s}$. Note that with X_t and X_s also the increment $X_t - X_s$ is a random variable. Furter, $(\Omega, \mathcal{A}, P, (X_t)_{t\in T})$ is a stochastic process with stochastically independent increments, i.e. for all $J = \{t_1, \dots, t_n\} \in \mathcal{H}(T)$ (where $\mathcal{H}(T) \coloneqq \{J | J \subset T, 0 < |J| < \infty\}$) with $n \geq 2$ and $t_1 < \cdots < t_n$ the random variables (increments) $X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}}$ are stochastically independent.

Note that, formally, the mathematical framework of stochastic processes requires an existence proof of the postulated probability measure P on $(\Omega, \mathcal{A})$, which can be achieved by showing the consistency of the set of marginal distributions defined above. We eschew this existence proof here and instead note some informal properties of Wiener processes required for the concept of stochastic integration.

Properties of Wiener processes

We first reformulate the Wiener process without reference to the underlying measure space and, from now on, denote the Wiener process by $W(t), t \in [0, T]$ (Hassler, 2007). Informally, $W(t)$ is a process with a starting value of 0, and independent, normally distributed, stationary increments. More formally, we can restate (1) – (4) in the definition of the Wiener process above as follows

(1) *Initial value* The initial value of the Wiener process is zero with probability 1, i.e. $P(W(0) = 0) = 1$.
(2) *Normality* The increments $W(t_1) - W(t_0), \dots, W(t_n) - W(t_{n-1})$, where $0 \leq t_0 \leq t_1 \leq \cdots \leq t_n$ are independent for $n \in \mathbb{N}$.

[2] In the following section, $|S|$ denotes the cardinality of a set S, not a determinant.

(3) *Stationarity and Independence* The increments $W(t) - W(s)$ are normally distributed with expectation 0 and variance equal to the temporal difference $s - t$ for $0 \leq s < t$, i.e. $W(t) - W(s) \sim N(W(t) - W(s); 0, t - s)$. Notably, the variance of the increments is stationary, i.e. it only depends on the difference $t - s$, but not the absolute values of s and t, and the joint distribution of the increments is multivariate normal with a diagonal covariance matrix.

The Wiener process is defined by the properties of its increments. Nevertheless, the Wiener process can, inituitively, be conceived as a stochastic function of the form $W: [0, T] \to \mathbb{R}, t \mapsto W(t)$ where (t) is distributed according to a normal distribution $N(W(t); 0, t)$. This is readily seen by considering the distribution of the increment $W(t) - W(0)$ for $t \in [0, T]$. According to (3) $W(t) - W(0) \sim N(W(t) - W(0); 0, t - 0)$ and according to (1) $W(0) = 0$ with probability 1. Based on the properties of normal distributions, we can thus infer that $\mathbb{E}\big(W(t)\big) = 0$ and $\mathbb{V}\big(W(t)\big) = t$.

Gaussian Processes

Equivalently, a Wiener process $W(t)$ can be conceived as a Gaussian process. A Gaussian process is a stochastic process $(X_t)_{t \in T}$ with a countable index set $T \coloneqq \{t_1, \dots, t_n\}$, $t_1 \leq t_2 \leq \cdots \leq t_n$, where $|T| = n, n \in \mathbb{N}$, which is multivariate normally distributed. In terms of Gaussian processes, the vector $W(t) \coloneqq \big(W(t_1), \dots, W(t_n)\big)^T$ is distributed according to $N(W(t);\ 0, \Sigma)$, where the entries of the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}, \succ \Sigma$ are given by $(\Sigma)_{ij} = \min(t_i, t_j)$ (for a proof, see e.g. (Hassler, 2007)). Note in particular, that the diagonal elements of this covariance matrix are increasing and that the off-diagonal elements are non-zero, i.e. modeling stochastic dependencies.

Stochastic integration

Stochastic integrals, in contradistinction to the integrals known from standard calculus and analysis, are random variables, implying that the values they assume display some "randomness". The aim of this section is to introduce the Ito integral of a stochastic process as the prerequisite for the framework of stochastic differential equations. To help with its introduction, the Ito integral is developed as the endpoint of a "stochastification" of the Riemann integral known from standard calculus. The development of this section is largely based on (Hassler, 2007), Sections 6 - 11.

Convergence in the quadratic mean

In this section, we lay the foundations for the definitions of stochastic integrals as the limits of their corresponding sum expressions, in analogy to the well-known quadrature expressions for Riemann integrals in deterministic analysis. To this end, we require a convergence notion for random variables and discuss the supporting interval partition used in the subsequent sections.

The convergence form required in the definition of stochastic integrals below is "convergence in the quadratic mean". A sequence of random variables $(X_n)_{n \in \mathbb{N}}$ is said to converge to a random variable X in the quadratic mean, if $\mathbb{E}((X_n - X)^2) \to 0$ for $n \to \infty$, which will be denoted as $X_n \xrightarrow{2} X$. Without proof, we state the following (Cauchy) convergence criterion: A sequence of random variables with $\mathbb{E}(X_n^2) < \infty$ converges in the quadratic mean, if for arbitrary n and m, $\mathbb{E}((X_m - X_n)^2) \to 0$ for $m, n \to \infty$, or, equivalently, if for arbitrary n and m, $\mathbb{E}(X_m X_n) \to c < \infty$ for for $m, n \to \infty, c \in \mathbb{R}$.

The concept of convergence in the quadratic mean can be demonstrated using a simple example: Consider the sequence $(X_n)_{n \in \mathbb{N}}$ of random variables defined by $X_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} Y_i$ for $n = 1,2, \dots$ where the $Y_i, i = 1, \dots, n$ are independent, normally distributed random variables with mean 0 and variance σ^2, i.e. $Y_i \sim N(Y_i;\ 0, \sigma^2)$ for

$i = 1, \dots, n$. Using the Cauchy convergence criterion above, it can be readily shown that this sequence converges in the quadratic mean (assuming $\mathbb{E}(X_n^2) < \infty$) as follows

$$\mathbb{E}(X_m X_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \cdot \frac{1}{m}\sum_{i=1}^{m} Y_i\right) = \frac{1}{mn}\sum_{i=1}^{\min(n,m)} \mathbb{E}(Y_i^2) = \sigma^2 \frac{\min(n,m)}{mn} \to 0 \tag{3.2}$$

for $n, m \to \infty$. In the above, the second equation follows, because the expectations of the products $Y_i \cdot Y_j$ for $i \neq j$ are 0, as the $Y_i, i = 1,2, \dots$ are assumed to be stochastically independent.

Interval partitions

In order to be able to define the integral of a function on the interval $[0, T] \subset \mathbb{R}$, the interval $[0, T]$ is partitioned into disjoint subintervals of the form

$$S_n([0,T]) \coloneqq [t_0, t_1] \cup [t_1, t_2] \cup \dots \cup [t_{n-1}, t_n] \tag{3.3}$$

where $0 = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = T$. It is assumed that this partition becomes more and more fine grained as n increases, in other words: $\max_{1 \le i \le n}(t_i - t_{i-1}) \to 0$ for $n \to \infty$. A typical example of a partition that fulfills this criterion is the equidistant partition of $[0, T]$, given by $t_i = \frac{iT}{n}$ $(i \in \mathbb{N}_n)$, for which one has

$$t_i - t_{i-1} = i\frac{T}{n} - (i-1)\frac{T}{n} = \frac{T}{n} \quad (i \in \mathbb{N}_n) \tag{3.4}$$

which obviously fulfills $\lim_{n\to\infty} \frac{T}{n} = 0$. We note that for any deterministic function $f: [0, T] \to \mathbb{R}$,

$$\begin{aligned} \sum_{i=1}^{n} f(t_i) - f(t_{i-1}) &= f(t_1) - f(t_0) + f(t_2) - f(t_1) + \cdots + f(t_n) - f(t_{n-1}) \\ &= f(t_n) - f(t_0) \\ &= f(T) - f(0) \end{aligned} \tag{3.5}$$

and thus, for the identity function, we have

$$\sum_{i=1}^{n} t_i - t_{i-1} = T - 0 = T \tag{3.6}$$

An arbitrary point of support in the interval $[t_{i-1}, t_i]$ will be denoted by t_i^*.

The stochastic Riemann integral

Informally, stochastic Riemann integrals are the usual Riemann integrals with a stochastic process being part of the function integrated over. If the stochastic process in question is a Wiener process, the stochastic Riemann integral is normally distributed with an expectation of zero and a closed form equation for its variance can be obtained. Formally, the general stochastic Riemann integral over the product of a deterministic function f and a general stochastic process $X(t)$ is defined as the limit of the Riemann sum

$$R_n = \sum_{i=1}^{n} f(t_i^*) X(t_i^*)(t_i - t_{i-1}) \tag{3.7}$$

If the limit of this sum convergences in the quadratic mean for $n \to \infty$ independently of the form of the partition $S_n([0, T])$, this limit is defined as the stochastic Riemann integral:

$$\int_0^T f(s)X(s)\, ds \coloneqq \lim_{n\to\infty} \sum_{i=1}^{n} f(t_i^*) X(t_i^*)(t_i - t_{i-1}) \tag{3.8}$$

We next consider stochastic Riemann integrals, for which the stochastic process is a Wiener process. For a deterministic function f, the stochastic Riemann integral of the product between f and a Wiener process $W(t)$ corresponds to a normal distribution with expectation

$$\mathbb{E}\left(\int_0^T f(s)W(s)\,ds\right) = 0 \tag{3.9}$$

and covariance

$$\mathbb{C}\left(\int_0^T f(s)W(s)\,ds, \int_0^T f(s)W(s)\,ds\right) = \int_0^T \int_0^T f(s)f(t)\min(s,t)\,dsdt \tag{3.10}$$

(for a proof of the above, see (Hassler, 2007)).

The Riemann-Stieltjes integral

Riemann-Stieltjes integrals are solutions for specific stochastic differential equations and are normally distributed random variables. The integrand of a Rieman-Stieltjes integral is a deterministic function f, however, this is integrated with respect to a Wiener process, which defines the Riemann-Stieltjes sum as follows

$$RS_n = \sum_{i=1}^n f(t_i^*)(W(t_i) - W(t_{i-1})) \tag{3.11}$$

As for the stochastic Riemann integral, if the limit of this sum for $n \to \infty$ exists in the quadratic mean sense independently of the form of the partition $S_n([0,T])$, this limit is defined as the Riemann-Stieltjes integral

$$\int_0^T f(s)\,dW(s) \coloneqq \lim_{n\to\infty} \sum_{i=1}^n f(t_i^*)(W(t_i) - W(t_{i-1})) \tag{3.12}$$

As a result from being the limit of a sum of normally distributed random the Riemann-Stieltjes integral is normally distributed with expectation

$$\mathbb{E}\left(\int_0^T f(s)\,dW(s)\right) = 0 \tag{3.13}$$

and covariance

$$\mathbb{C}\left(\int_0^T f(s)\,dW(s), \int_0^T f(s)\,dW(s)\right) = \int_0^T f^2(s)\,ds \tag{3.14}$$

(for a proof of the above, see (Hassler, 2007)). To see that for the Riemann-Stieltjes integral of the constant unity function

$$\int_{t_1}^{t_2} dW(s) = W(t_2) - W(t_1) \tag{3.15}$$

we consider its relation to the concept of partial integration as known from the calculus of standard Riemann integrals. Recall that, for two deterministic integrable functions $f\colon [0,T] \to \mathbb{R}$ and $g\colon [0,T] \to \mathbb{R}$, the following rule of integration holds

$$\int_0^T f(t)g'(t)dt = f(t)g(t)|_0^T - \int_0^T f'(t)g(t)dt \tag{3.16}$$

For the Riemann-Stieltjes integral, one has (changing the upper integral bound from T to $t \in [0,T]$, and the variable of integration from t to s, for a proof see (Hassler, 2007))

$$\int_0^t f(s)dW(s) = f(s)W(s)|_0^t - \int_0^t W(s)df(s) \tag{3.17}$$

which can be reformulated as

$$\int_0^t f(s)dW(s) = f(t)W(t) - \int_0^t W(s)f'(s)ds \tag{3.18}$$

Finally, we consider the special case of the Riemann-Stieltjes integral of the unity function $f(t) \coloneqq 1$. Substitution into the equation above yields

$$\int_0^t dW(s) = W(t) - W(0) - \int_0^t W(s) \cdot 0 \, ds = W(t) \tag{3.19}$$

More generally, on the interval $[t_1, t_2] \subset [0, T]$, one thus has

$$\int_{t_1}^{t_2} dW(s) = W(s)|_{t_1}^{t_2} - \int_{t_1}^{t_2} W(s) \cdot 0 \, ds = W(t_2) - W(t_1) \tag{3.20}$$

The Ito integral

The Ito integral provides the general basis for stochastic differential equations as discussed below. Here, we briefly discuss its definition and some of its properties. The general Ito integral of a stochastic process $X(t)$ with respect to a Wiener process $W(t)$ is defined as limit of the Ito sum

$$I_n \coloneqq \sum_{i=1}^n X(t_{i-1})((W(t_i) - W(t_{i-1})) \tag{3.21}$$

where, importantly (and in contradistinction to other stochastic integrals, as e.g., Stratonivich integral) the point t_{i-1}, i.e., the left border of the partition interval $[t_{i-1}, t_i]$ serves as supporting point. If $X(t)$ is a stochastic process with finite variance, and if $X(t)$ is independent of the future of the stochastic process, the Ito sum converges to a well-defined limit in the quadratic means sense and is defined as the Ito integral

$$\int_0^T X(t) \, dW(t) \coloneqq \lim_{n\to\infty} \sum_{i=1}^n X(t_{i-1})((W(t_i) - W(t_{i-1})) \tag{3.22}$$

The Ito integral is a random variable and its first two central moments can be shown to be given by (for a proof see (Hassler, 2007))

$$\mathbb{E}\left(\int_0^T X(t) \, dW(t)\right) = 0 \tag{3.23}$$

and

$$\mathbb{C}\left(\int_0^T X(t) \, dW(t), \int_0^T X(t) \, dW(t)\right) = \int_0^T \mathbb{E}\big(X^2(t)\big) \, dt \tag{3.24}$$

In the degenerate case that the stochastic process $X(t)$ corresponds to the unity function, i.e., that $X(t) \coloneqq 1$ with probability $P\big(X(t)\big) = 1$, the Ito integral evaluates to the Riemann-Stieltjes integral discussed in the previous section, which in turn evaluates to the increment of a Wiener process (as seen above, again changing the upper integral bound from T to $t \in [0, T]$, and the variable of integration from t to s):

$$\int_0^t X(s) \, dW(s) = \int_0^t 1 \, dW(s) = W(t) \tag{3.25}$$

Stochastic Differential Equations

Diffusions and Stochastic Differential Equations (SDEs)

A diffusion is a stochastic process comprising the sum of a stochastic Riemann integral and an Ito integral in the form

$$X(t) = X(0) + \int_0^t \mu(X(s), s) \, ds + \int_0^t \sigma(X(s), s) \, dW(s) \tag{3.26}$$

where $X(0)$ indicates the starting point of the stochastic process X. Here, the functions μ and σ are referred to as a drift function and volatility functions, respectively, and are functions of time and the diffusion process itself. The first integral term, a stochastic Riemann integral, models a deterministic drift of the process X, while the second integral term, an Ito integral, models a random component. Taking the derivative with respect to time on both sides of the diffusion equation above yields its differential form

$$\frac{d}{dt}X(t) = \frac{d}{dt}\left(X(0) + \int_0^t \mu(X(s), s)\, ds + \int_0^t \sigma(X(s), s)\, dW(s)\right) \tag{3.27}$$

which usually denoted as

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \tag{3.28}$$

Informally, the equation above states, that the change of the process X within an "infinitesimal" small time interval at time point t is given as the sum of (1) the product of a deterministic function μ (depending on the state of the process X at time point t and time itself) with the "infinitesimal" length of the time interval dt and (2) the product of a deterministic function σ (also depending on the state of the process X at time t and time itself) with the "infinitesimal" small increment of a Wiener process at time t, which is randomly distributed. The classification of stochastic differential equations (SDEs), as well as the study of existence and uniqueness properties of their solutions, forms an integral part of stochastic analysis. The interested reader is referred to (Kloeden & Platen, 1999) in depth analytical and numerical treatments of SDEs. Here, we briefly introduce the set of narrow-sense linear SDEs with the aim of introducing the Langevin equation as provided in equation (4).

Autonomous Narrow-sense linear SDEs and the Langevin equation

An SDE of the type

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \tag{3.29}$$

is called a linear SDE, if the functions $\mu\colon \mathbb{R} \times [0, T] \to \mathbb{R}$ and $\sigma\colon \mathbb{R} \times [0, T] \to \mathbb{R}$ take the form

$$\mu(X(t), t) \coloneqq \mu_1(t)X(t) + \mu_2(t) \text{ and } \sigma(X(t), t) \coloneqq \sigma_1(t)X(t) + \sigma_2(t) \tag{3.30}$$

for coefficients $\mu_1, \mu_2, \sigma_1, \sigma_2\colon [0, T] \to \mathbb{R}$. If all coefficients are constants, the linear SDE is called autonomous. Further, if $\mu_2(t) = \sigma_2(t) \coloneqq 0$, the linear SDE is referred to as homogenous. Finally, the linear SDE is referred to as linear in the narrow sense, if $\sigma_1(t) \coloneqq 0$, or in other words, the noise term is additive. The autonomous narrow-sense linear SDE thus has the general form

$$dX(t) = (\mu_1 X(t) + \mu_2)dt + \sigma_2 dW(t) \tag{3.31}$$

Setting $\mu_1 \coloneqq \alpha, \mu_2 \coloneqq 0$ and $\sigma_2 \coloneqq \sigma$, one obtains the "latent linear diffusion process" of equation (ref. equation 4 of the main tutorial), commonly known as the Langevin equation:

$$dX(t) = \alpha X(t)dt + \sigma dW(t)\ (\alpha < 0, \sigma > 0, t \in [0, T]) \tag{3.32}$$

The general solution of the Langevin equation is given by (for a proof see (Hassler, 2007))

$$X(t) = \exp(\alpha t)X(0) + \exp(\alpha t)X(0) \int_0^t \sigma \exp(s)\, dW(s) \tag{3.33}$$

which specifies a homogenous Gaussian process with expectation and variance

$$\mathbb{E}\big(X(t)\big) = \exp(\alpha t)\, \mathbb{E}(X(0)) \tag{3.34}$$

and

$$\mathbb{V}\big(X(t)\big) = \exp(\alpha t)\,\mathbb{V}\big(X(0)\big) + \frac{\sigma^2}{2\alpha}(1 - \exp(\alpha t)) \tag{3.35}$$

Euler Maruyama discretization

A commonly employed discretization technique for SDEs is the Euler-Maruyama method (Hassler, 2007; Kloeden & Platen, 1999), which defines a recursive scheme for the numerical evaluation of SDEs on a time interval $[0, T] \subset \mathbb{R}$. The Euler-Maruyama method is based at a set of discrete time points $t_i \in [0, T],\ i = 0,1, \dots, n$ with

$$0 = t_0 < t_1 = \frac{T}{n} < \cdots < t_{n-1} < t_n = T$$

which represent the discretization of the observation time interval

$$[0, T] = \cup_{i=1}^{n} \left[\frac{i-1}{n}T, \frac{i}{n}T\right[\tag{3.36}$$

into n equisized bins of length $\Delta t \coloneqq \frac{T}{n}$, and which in the context of the tutorial may be conceived as being provided by the sampling rate of the measurement or observation process. The first step towards the Euler-Maruyama approximation of (3.32) to rewrite it in its integral form given by

$$X(t) = X(0) + \int_0^t \alpha X(s)ds + \int_0^t \sigma dW(s) \qquad t \in [0, T] \subset \mathbb{R} \tag{3.37}$$

In (3.37) the first integral term denotes a stochastic Riemann integral and the second term denotes a Riemann-Stieltjes integral. Considering this integral form on the subinterval $[t_{i-1}, t_i[\subset [0, T]\ \ (i = 1, \dots, n)$ then yields

$$X(t_i) = X(t_{i-1}) + \int_{t_{i-1}}^{t_i} \alpha X(s)ds + \int_{t_{i-1}}^{t_i} \sigma dW(s) \tag{3.38}$$

By approximating the value of $X(s)$ for $s \in [t_{i-1}, t_i]$ by its left endpoint value, that is, by setting $X(s) \coloneqq X(t_{i-1})$ for all $s \in [t_{i-1}, t_i]$, one obtains the Euler-Maruyama approximation of the stochastic differential equation

$$X(t_i) = X(t_{i-1}) + \int_{t_{i-1}}^{t_i} \alpha X(t_{i-1})ds + \int_{t_{i-1}}^{t_i} \sigma dW(s) \tag{3.39}$$

$$= X(t_{i-1}) + \alpha X(t_{i-1}) \int_{t_{i-1}}^{t_i} ds + \sigma \int_{t_{i-1}}^{t_i} dW(s) \tag{3.40}$$

for $i = 1, \dots, n$. Riemann integration of the first integral term yields

$$\int_{t_{i-1}}^{t_i} ds = s|_{t_{i-1}}^{t_i} = t_i - t_{i-1} = \frac{T}{n} = \Delta t \tag{3.41}$$

and Rieman-Stieltjes integration of the second integral yields

$$\int_{t_{i-1}}^{t_i} dW(s) = W(t_i) - W(t_{i-1}) \tag{3.42}$$

From the definition of Wiener processes we note that that

$$W(t_i) - W(t_{i-1}) = W\left(\frac{iT}{n}\right) - W\left(\frac{(i-1)T}{n}\right) \tag{3.45}$$

is normally distributed with expectation parameter 0 and variance parameter Δt.To rewrite

$$dX(t) = \alpha X(t)dt + \sigma dW(t) \tag{3.46}$$

in the familiar autoregressive process of order 1 (AR(1)) form

$$x_t = a x_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim N(\varepsilon, ; 0, \sigma_x^2) \tag{3.47}$$

we thus start by considering

$$X(t_i) = X(t_{i-1}) + \alpha X(t_{i-1})\Delta t + \sigma\big(W(t_i) - W(t_{i-1})\big) \tag{3.48}$$

and change the notation by setting

$$x_{t_i} \coloneqq X(t_i),\ w_{t_i} \coloneqq W(t_i) \text{ and } \varepsilon_{t_i} \coloneqq w_{t_i} - w_{t_{i-1}} \tag{3.49}$$

resulting in

$$x_{t_i} = x_{t_{i-1}} + \alpha x_{t_{i-1}}\Delta t + \sigma\varepsilon_{t_i} \text{ where } \varepsilon_{t_i} \sim N(\varepsilon_{t_i}; 0, \Delta t) \tag{3.50}$$

Next replace the discrete real time points $0 = t_0, \dots, t_n = T$ by integer indices $t = 0, 1, 2 \dots, T \coloneqq n$ resulting in

$$x_t = x_{t-1} + \alpha x_{t-1}\Delta t + \sigma\varepsilon_t \text{ where } \varepsilon_t \sim N(\varepsilon_t; 0, \Delta t) \tag{3.51}$$

Using $\mathbb{V}(aX) = a^2\mathbb{V}(X)$, we can rewrite the above as

$$x_t = (1 + \alpha\Delta t)x_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(\varepsilon_t; 0, \sigma^2\Delta t) \tag{3.52}$$

Finally, we define

$$a \coloneqq (1 + \alpha\Delta t) \text{ and } \sigma_x^2 \coloneqq \sigma^2\Delta t$$

and obtain the AR(1) form

$$x_t = ax_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(\varepsilon_t; 0, \sigma_x^2) \tag{3.53}$$

To transfer probabilistic statements over LGSSM parameters to probabilistic statements over SDE parameters, we use the probability density function transformation theorem in Supplement Section 9. To this end, it is helpful to note that the transformation mappings from LGSSM to SDE parameters, and their inverses are given as

$$T_1 \colon \mathbb{R} \to \mathbb{R}, a \mapsto T_1(a) = \frac{1}{\Delta t}(a - 1) \tag{3.54}$$

with

$$T_1^{-1} \colon \mathbb{R} \to \mathbb{R}, \alpha \mapsto T_1^{-1}(\alpha) = (1 + \alpha\Delta t) \tag{3.55}$$

and, because we will derive probabilistic statements on the inverse λ_x of σ_x^2 using $\lambda_x = \frac{1}{\sigma^2\Delta t}$

$$T_2 \colon \mathbb{R}_+ \to \mathbb{R}_+, \lambda_x \mapsto T_2(\lambda_x) = \frac{\Delta t}{\sqrt{\lambda_x}} \tag{3.56}$$

with inverse

$$T_2^{-1} \colon \mathbb{R}_+ \to \mathbb{R}_+, \sigma \mapsto T_2^{-1}(\sigma) \coloneqq \left(\frac{\Delta t}{\sigma}\right)^2 \tag{3.57}$$

4 An Introduction to Variational Calculus

The basic problem of variational calculus is the optimization, i.e. minimization or maximization, of functionals with respect to their arguments, usually functions on real vector spaces. In other words, given a functional $\mathcal{F}$ defined on a function space V, the aim is to find an element y of V for which $\mathcal{F}$ assumes an extremal value. This basic problem can be understood in analogy to the maximization/minimization of a function defined on a standard vector space, such as $\mathbb{R}$ or $\mathbb{R}^n$. In fact, the basic approach of solving the variational optimization problem shows many intuitive similarities with approaches from ordinary calculus (Figure 3, panels A and B). In ordinary calculus, the argument x that maximizes a given function f is usually found by first identifying the stationary points of f, i.e. those arguments x, for which the derivative or gradient of f vanishes. This is the necessary condition for an extremal value: If f has a maximum or a minimum in x, then its gradient in x is equal to zero. However, the condition of a vanishing gradient is not sufficient for a (local) extremal value: The gradient may also vanish for an inflection point or for constant parts of f. Similar considerations apply for the calculus of variations. The formal treatment of necessary and sufficient conditions of functionals is a central topic of functional analysis, just as real analysis represents a formal treatment of ordinary calculus. In the current tutorial, functional analysis is eschewed and only the necessary condition for extremals of functionals will be considered. This approach is somewhat justified by the applied character of this tutorial and the intuition that the functionals considered are usually convex (for convex functions, the necessary condition for an extremal is also sufficient (Boyd & Vandenberghe, 2004)). The basic problem of variational calculus may thus be formalized as follows: Let V be a vector space of functions, for example, the space of all real-valued differentiable functions on the real line $V \coloneqq \{y \in \mathcal{C}^1(\mathbb{R})\}$. Further, let $\mathcal{F}$ be a real-valued functional on V, i.e., let $\mathcal{F}$ be a mapping that takes an element y of V as input and maps it onto the real line, formally $\mathcal{F}: V \to \mathbb{R}$. Then, the basic problem of variational calculus may be stated as finding $y^* \in V$ for which $\mathcal{F}$ assumes a minimum (or $-\mathcal{F}$ assumes a maximum):

$$y^* \coloneqq \arg\min_{y \in V} \mathcal{F}(y) = \arg\max_{y \in V} \big(-\mathcal{F}(y)\big) \tag{4.1}$$

A necessary condition for such an extremal argument (and, by appealing to the intuition from ordinary calculus, an equation of the type $f'(x) = 0$ which can be solved for the respective argument x) can be established by introducing the so-called "Gateaux derivative" of $\mathcal{F}$: Let y^* be a minimum of $\mathcal{F}$ and let $v \in V$ be another function in V. Because V is a function space, $y^* + \varepsilon v$ for $\varepsilon \in \mathbb{R}$ is also an element of V. Moreover, because y^* is a minimum of $\mathcal{F}$

$$\mathcal{F}(y^* + \varepsilon v) \geq \mathcal{F}(y^*) \tag{4.2}$$

This consideration allows for defining a function of ε as follows:

$$h_{y^*,v}: \mathbb{R} \to \mathbb{R},\ \varepsilon \mapsto h_{y^*,v}(\varepsilon) \coloneqq \mathcal{F}(y^* + \varepsilon v) \tag{4.3}$$

By definition, this function has a minimum in $\varepsilon = 0$ as $\mathcal{F}(y^* + 0 \cdot v) = \mathcal{F}(y^*) \leq \mathcal{F}(y^* + \varepsilon v)$. Thus, if $h_{y^*,v}$ is differentiable, its derivative with respect to ε vanishes in $\varepsilon = 0$, in other words,

$$\frac{d}{d\varepsilon} h_{y^*,v}(\varepsilon)|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{\mathcal{F}(y^* + \varepsilon v) - \mathcal{F}(y^*)}{\varepsilon} = 0 \tag{4.4}$$

This derivative of $h_{y^*,v}$ in $\varepsilon = 0$ (if defined) is referred to as the Gateaux derivative of $\mathcal{F}$ and is usually denoted as $\delta\mathcal{F}(y^*, v)$. One thus obtains

$$\delta\mathcal{F}(y^*, v) \coloneqq \lim_{\varepsilon \to 0} \frac{\mathcal{F}(y^* + \varepsilon v) - \mathcal{F}(y^*)}{\varepsilon} = 0 \Leftrightarrow \frac{d}{d\varepsilon} h_{y^*,v}(\varepsilon)|_{\varepsilon=0} = 0 \tag{4.5}$$

A necessary condition for an extremal value of the functional $\mathcal{F}$ can thus be expressed as follows: Let y^* be an extremal value of $\mathcal{F}$, i.e., $y^* \coloneqq \arg\min_{y \in V} \mathcal{F}(y)$. Then it follows that

$$\delta\mathcal{F}(y^*, v) = 0 \tag{4.6}$$

for all $v \in V$ given that $\delta\mathcal{F}(y^*, v)$ exists. This condition thus provides an approach for determining an extremal argument of a functional: after determining $\delta\mathcal{F}(y, v)$ for a given function $\mathcal{F}$, set $\delta\mathcal{F}(y, v) = 0$ and solve for y. This approach yields the stationary point(s) of $\mathcal{F}$, which can then be investigated with respect to their extremal properties. Further, in the case of a convex function $\mathcal{F}$, the stationary points already constitute the minimum/maximum argument(s).

4.1 The Variational Problem with Fixed Endpoints

Above, a necessary condition for an extremal argument y^* of a functional $\mathcal{F}$ was introduced within a very general context. Here, this general approach is made concrete for a specific class of functionals (subsuming, in a sense, the variational free energy considered in Section 4 of the main tutorial) and is demonstrated by means of an example. Specifically, by limiting the function space to real-valued functions on an interval $[x_0, x_1]$ and specifying the class of functionals as the integrals of the so-called Lagrange functions on this interval, one arrives at the Euler-Lagrange equation for functional optimization. As will be shown below, the Euler-Lagrange equation defines the extremal argument y^* of a functional $\mathcal{F}$ by means of a second-order partial differential equation, which can be solved using methods from ordinary calculus.

The variational fixed endpoint may be stated as follows: let $(x_0, y_0), (x_1, y_1)$ be elements of $\mathbb{R}^2$ and V the vector space of all twice-differentiable functions y on (x_0, x_1) which fulfill the endpoint conditions $y(x_0) = y_0$ and $y(x_1) = y_1$, i.e.,

$$V \coloneqq \{y \in \mathcal{C}^2[x_0, x_1] | y(x_0) = y_0, y(x_1) = y_1\} \tag{4.7}$$

Further, let L be a twice partially differentiable, real-valued function defined on the set $\mathbb{R} \times \mathbb{R} \times [x_0, x_1]$, referred to as the Lagrange function:

$$L: \mathbb{R} \times \mathbb{R} \times [x_0, x_1] \to \mathbb{R} \tag{4.8}$$

Then, the variational fixed endpoint problem consist of finding a $y^* \in V$ such that the functional $\mathcal{F}$ defined by

$$\mathcal{F}: V \to \mathbb{R}, y \mapsto \mathcal{F}(y) \coloneqq \int_{x_0}^{x_1} L(y(x), y'(x), x)\, dx \tag{4.9}$$

becomes extremal, in other words, finding

$$y^* \coloneqq \arg\min_{y \in V} \int_{x_0}^{x_1} L(y, y', x)\, dx \tag{4.10}$$

Note that, y and y' refer to both the functions $y, y' \in V$ and the real coordinates $y(x), y'(x) \in \mathbb{R}$ of the Lagrange function L for $x \in \mathbb{R}$.

Specifying the Gateaux derivative of $\mathcal{F}$ as in (4.5) for functions $v \in V$ with $v(x_0) = v(x_1) = 0$ then yields the following necessary condition for an extremal value in the current fixed endpoint case

$$\delta\mathcal{F}(y^*, v) = \frac{d}{d\varepsilon} \int_{x_0}^{x_1} L(y^* + \varepsilon v, y^{*\prime} + \varepsilon v', x)\, dx|_{\varepsilon=0} = 0 \tag{4.11}$$

It can be shown that the derivative of the integral in (4.11) may reformulated such that the necessary condition above may equivalently be expressed as (see below for details)[3]

[3] For simplicity of notation, the asterisk denoting the extremal argument is suppressed in the following equations, which also emphasizes the notion of y being a variable which is solved for.

$$\delta\mathcal{F}(y,v) = 0 \Leftrightarrow \frac{\partial}{\partial y}L(y,y',x) - \frac{d}{dx}\frac{\partial}{\partial y'}L(y,y',x) = 0 \tag{4.12}$$

The equation on the right-hand side of the expression above is referred to as the Euler-Lagrange equation and represents an (implicit) second-order partial differential equations for y. This can be seen by evaluating the total differential of its second term resulting in (see below for details)

$$\frac{\partial}{\partial y}L(y,y',x) = \frac{d}{dx}\frac{\partial}{\partial y'}L(y,y',x) \tag{4.13}$$

$$\Leftrightarrow \frac{\partial}{\partial y}L(y,y',x) = \frac{\partial^2}{\partial y \partial y'}L(y,y',x)y' + \frac{\partial^2}{\partial y' \partial y'}L(y,y',x)y'' + \frac{\partial^2}{\partial x \partial y'}L(y,y',x)$$

This implicit partial differential equation for y may be made explicit by dividing the second partial derivative of the Lagrange function with respect to y' and evaluated using standard methods from ordinary calculus. The resulting $y \in V$ is a stationary point of $\mathcal{F}$ and a candidate for an extremal argument. Again, if $\mathcal{F}$ is convex, it is also already known to be an extremal argument. In the following example, the calculus of partial differential equations is eschewed, as the functional considered is an explicit function of y' only.

○ **Derivation of equations (4.12) and (4.13)**

In this section, we derive the Euler-Lagrange equation

$$\frac{\partial}{\partial y}L(y,y',x) - \frac{d}{dx}\frac{\partial}{\partial y'}L(y,y',x) = 0 \tag{4.14}$$

based on the necessary condition for a minimum of

$$\mathcal{F}(y) \coloneqq \int_0^l L(y(x),y'(x),x)\,dx \tag{4.15}$$

given by

$$\delta\mathcal{F}(y,v) = \frac{d}{d\varepsilon}\int_0^l L(y+\varepsilon v, y'+\varepsilon v', x)\,dx|_{\varepsilon=0} = 0 \tag{4.16}$$

To this end, we first note the definition of the total differential for a function $f\colon \mathbb{R}^3 \to \mathbb{R}$ and $y = (y,y',x)^T, v = (v,v',0)^T, \varepsilon \in \mathbb{R}$ is given by

$$\frac{d}{d\varepsilon}f(y+\varepsilon v)|_{\varepsilon=0} = \sum_{i=1}^n \frac{\partial}{\partial y_i}f(y)\cdot v_i \tag{4.17}$$

Because the Lagrange function L is differentiable by definition, we thus obtain

$$\frac{d}{d\varepsilon}\int_0^l L(y+\varepsilon v, y'+\varepsilon v', x)\,dx|_{\varepsilon=0} = \int_0^l \frac{d}{d\varepsilon}(L(y+\varepsilon v, y'+\varepsilon v', x))|_{\varepsilon=0}\,dx \tag{4.18}$$

$$= \int_0^l \frac{\partial}{\partial y}L(y,y',x)\cdot v + \frac{\partial}{\partial y'}L(y,y',x)\cdot v' + \frac{\partial}{\partial y'}L(y,y',x)\cdot 0\,dx$$

$$= \int_0^l \frac{\partial}{\partial y}L(y,y',x)v + \frac{\partial}{\partial y'}L(y,y',x)v'\,dx$$

$$= \int_0^l \frac{\partial}{\partial y}L(y,y',x)v\,dx + \int_0^l \frac{\partial}{\partial y'}L(y,y',x)v'\,dx$$

The second integral term is next integrated using partial integration. In general we have

$$\int_a^b f'g = fg|_a^b - \int_a^b fg' \tag{4.19}$$

Here, we have

$$a = 0,\ b = l,\ g = \frac{\partial}{\partial y'}L,\ g' = \frac{d}{dx}\frac{\partial}{\partial y'}L,\ f = v,\ f' = v' \tag{4.20}$$

and thus obtain

$$\int_0^l \frac{\partial}{\partial y'}L \cdot v'\,dx = \frac{\partial}{\partial y'}L \cdot v|_0^l - \int_0^l v \cdot \frac{d}{dx}\frac{\partial}{\partial y'}L \tag{4.21}$$

Written in full, the above yields

$$\int_0^l \frac{\partial}{\partial y'(x)}L(y(x), y'(x), x)v'\,dx \tag{4.22}$$

$$= \frac{\partial}{\partial y'(x)}L(y(x), y'(x), x)v(x)|_0^l - \int_0^l \frac{d}{dx}\frac{\partial}{\partial y'(x)}L\,(y(x), y'(x), x)v(x)dx$$

$$= \frac{\partial}{\partial y'(x)}L(y(l), y'(l), l)v(l) - \frac{\partial}{\partial y'(x)}L(y(0), y'(0),0)v(0) - \int_0^l \frac{d}{dx}\frac{\partial}{\partial y'(x)}L\,(y(x), y'(x), x)v(x)dx$$

Substitution into the equation above results in

$$\delta\mathcal{F}(y, v) = 0 \tag{4.23}$$

$$\Leftrightarrow \int_0^l \frac{\partial}{\partial y(x)}L(y(x), y'(x), x)v(x)\,dx + \frac{\partial}{\partial y'(l)}L(y(l), y'(l), l)v(l) - \frac{\partial}{\partial y'(0)}L(y(0), y'(0), 0)v(0) - \int_0^l \frac{d}{dx}\frac{\partial}{\partial y'(x)}L\,(y(x), y'(x), x)v(x)dx = 0$$

$$\Leftrightarrow \int_0^l \left(\frac{\partial}{\partial y}L(y(x), y'(x), x) - \frac{d}{dx}\frac{\partial}{\partial y'(x)}L(y(x), y'(x), x)\right)v(x)\,dx + \frac{\partial}{\partial y'(l)}L(y(l), y'(l), l)v(l) - \frac{\partial}{\partial y'(l)}L(y(0), y'(0), 0)v(0) = 0$$

As we are considering only those v for which $v(0) = v(l) = 0$, because the function $y + \varepsilon v$ is supposed to fulfill $y(0) + \varepsilon v(0) = y(0) = y_0$ and $y(l) + \varepsilon v(l) = y(l) = y_l$, we have

$$\delta\mathcal{F}(y, v) = 0 \Leftrightarrow \int_0^l \left(\frac{\partial}{\partial y(x)}L(y(x), y'(x), x) - \frac{d}{dx}\frac{\partial}{\partial y'(x)}L(y(x), y'(x), x)\right)v(x)\,dx = 0 \tag{4.24}$$

We now make use of the following Lagrange-Lemma, which we state without proof:

<u>*Lagrange-Lemma*</u>

Let f be a continuous real-valued function on the interval $[0, l]$ and let

$$\int_0^l f(x)v(x)\,dx = 0$$

For all continuously differentiable, real-valued functions$v$ that fulfill $v(0) = v'(0) = v(l) = v'(l) = 0$. Then it follows that $f(x) \equiv 0$ on $[0, l]$.

With this Lemma, we thus obtain the Euler-Lagrange equation

$$\delta\mathcal{F}(y, v) = 0 \Leftrightarrow \frac{\partial}{\partial y(x)}L(y(x), y'(x), x) - \frac{d}{dx}\frac{\partial}{\partial y'(x)}L(y(x), y'(x), x) = 0 \tag{4.25}$$

Finally, we note that the Euler-Lagrange equation constitutes a second order partial differential equation for the function y, because writing out the differential of the second term on its left hand side yields

$$\frac{d}{dx}\frac{\partial}{\partial y'(x)}L(y(x), y'(x), x) = \frac{\partial^2}{\partial y(x)\partial y'(x)}L(y(x), y'(x), x)y'(x) + \frac{\partial^2}{\partial y'(x)\partial y'(x)}L(y(x), y'(x), x)y''(x)) + \frac{\partial^2}{\partial y'(x)\partial x}L(y(x), y'(x), x)$$

(4.26)

Written in full, the Euler-Lagrange equation in its general form thus corresponds to

$$\frac{\partial}{\partial y(x)}L(y(x),y'(x),x)-\frac{\partial^2}{\partial y(x)\partial y'(x)}L(y(x),y'(x),x)y'(x)-\frac{\partial^2}{\partial y'(x)\partial y'(x)}L(y(x),y'(x),x)y''(x)-\frac{\partial^2}{\partial y'(x)\partial x}L(y(x),y'(x),x)=0 \tag{4.27}$$

□

Example: The Shortest Distance Between Two Points in $\mathbb{R}^2$

As an example, we consider the following problem: Let $(a,\alpha),(b,\beta)\in\mathbb{R}^2, a<b$. The vector space V may be defined as

$$V:=\{y\in\mathcal{C}^1[a,b]|y(a)=\alpha,y(b)=\beta\} \tag{4.28}$$

and the aim is to find a $y\in V$ for which the functional

$$\mathcal{F}:V\to\mathbb{R},y\mapsto\mathcal{F}(y):=\int_a^b\sqrt{1+y'(x)^2}\,dx \tag{4.29}$$

becomes minimal. The specific form of the functional in this example is motivated by the fact that, for parameterized curves in $\mathbb{R}^2$ (i.e., functions y of the type $f:\mathbb{R}\to\mathbb{R}^2$), the arc length is defined as

$$l:V\to\mathbb{R},l\mapsto l(f):=\int_a^b\|f'(x)\|_2\,dx \tag{4.30}$$

and the function y considered here represents the "function form" of such a curve, i.e., $f(x)=\big(x,y(x)\big)^T$ for which $f'(x)$ evaluates to $\big(1,y'(x)\big)^T$. In the special case of the example considered here the Lagrange function is given by

$$L:\mathbb{R}\to\mathbb{R},y'\mapsto L(y'):=\sqrt{1+y'^2} \tag{4.31}$$

L is thus not explicitly dependent on y and x, the partial derivatives of the Lagrange function with respect to y and x evaluate to zero, and the Euler-Lagrange equation simplifies to

$$\frac{\partial}{\partial y}L(y')=\frac{d}{dx}\frac{\partial}{\partial y'}L(y')\Leftrightarrow\frac{d}{dx}\frac{\partial}{\partial y'}L(y')=0 \tag{4.32}$$

The right-hand side of the equivalence relation (4.32) states that the total derivative of $\frac{\partial}{\partial y'}L(y')$ is zero and the function $\frac{\partial}{\partial y'}L(y')$ thus constant. One hence obtains an ordinary differential equation for y given by

$$\frac{\partial}{\partial y'}L(y')=0\Rightarrow y'(x)=\sqrt{\frac{c^2}{1-c^2}} \tag{4.33}$$

for a given constant $c\in\mathbb{R}$ (see below for details). As the derivative of y is thus constant, y must be a linear-affine function of the type

$$y(x)=sx+r,s,r\in\mathbb{R} \tag{4.34}$$

Based on the fixed endpoints $y(a)=\alpha,y(b)=\beta$, the constants s and r may be evaluated, resulting in the solution

$$y^*(t):=\arg\min_{y\in V}\int_a^b\sqrt{1+y'(x)^2}\,dx=\frac{\beta-\alpha}{b-a}(x-a)+\alpha \tag{4.35}$$

The application of the variational method hence verifies the intuition that the functional form of the curve representing the shortest distance between two points in the plane is a straight line (Figure 2 panels C and D).

○ **Derivation of equation (4.33)**

We first evaluate the partial derivative $\frac{\partial}{\partial y'}L(y')$ for $L(y') \coloneqq \sqrt{1+y'^2}$:

$$\frac{\partial}{\partial y'}L(y') = \frac{\partial}{\partial y'}(1+y'^2)^{\frac{1}{2}} = \frac{1}{2}(1+y'^2)^{-\frac{1}{2}} \cdot 2 \cdot y' = \frac{y'}{\sqrt{1+y'^2}} \tag{4.36}$$

The Euler Lagrange Equation

$$\frac{d}{dx}\left(\frac{y\prime}{\sqrt{1+y'^2}}\right) = 0 \tag{4.37}$$

thus implies that

$$\frac{y'}{\sqrt{1+y'^2}} = c \tag{4.38}$$

for a real constant $c \in \mathbb{R}$. Reformulating then leads to the insight that the derivative of y w.r.t. x is constant:

$$y'^2 = c^2(1+y'^2) \Leftrightarrow y'^2 = c^2 + c^2y'^2 \Leftrightarrow y'^2 - c^2y'^2 = c^2 \Leftrightarrow y'^2(1-c^2) = c^2 \Leftrightarrow y' = \sqrt{\frac{c^2}{1-c^2}} \tag{4.39}$$

The Isoperimetric Variational Problem

In addition to the conditions of fixed endpoints, the extremal function maximizing/minimizing a given functional in the context of isoperimetric problems have to fulfill additional integral conditions. In general, the isoperimetric variational problem may be stated as follows: let $(x_0, y_0), (x_1, y_1)$ be elements of $\mathbb{R}^2$ and V the vector space of all twice-differentiable functions y on (x_0, x_1) which fulfill the endpoint conditions $y(x_0) = y_0$ and $y(x_1) = y_1$. Further, let L and G be twice partially differentiable, real-valued functions defined on the set $\mathbb{R} \times \mathbb{R} \times [x_0, x_1]$. Then, the isoperimetric variational problem comprises finding a $y^* \in V$ such that the functional $\mathcal{F}$ defined by

$$\mathcal{F}: V \to \mathbb{R}, y \mapsto \mathcal{F}(y) \coloneqq \int_{x_0}^{x_1} L(y(x), y'(x), x)\, dx \tag{4.40}$$

becomes extremal, and the integral side-condition

$$I: V \to \mathbb{R}, y \mapsto I(y) \coloneqq \int_{x_0}^{x_1} G(y(x), y'(x), x)\, dx = 0 \tag{4.41}$$

The formal development of an approach for the solution of the isoperimetric variational problem proving the existence of Lagrangian multipliers falls into the realm of functional analysis and is eschewed here. Nevertheless, assuming the existence of both an extremal argument y^* and suitable Lagrangian multipliers λ, the Lagrange Lemma introduced in below provides the following recipe for solving an isoperimetric variational problem: (1) consider the extended Lagrange function $F: \mathbb{R} \times \mathbb{R} \times [x_0, x_1] \to \mathbb{R}$, defined by

$$F(y(x), y'(x), x) \coloneqq L(y(x), y'(x), x) + \lambda G(y(x), y'(x), x) \tag{4.42}$$

(2) instead of minimizing $\mathcal{F}$, minimize the extended functional

$$\tilde{\mathcal{F}} \coloneqq \int_{x_0}^{x_1} F(y(x), y'(x), x)\, dx \tag{4.43}$$

using the fixed endpoint approach developed in Section 3.2 (usually by means of the corresponding Euler-Lagrange equations) and (3) simultaneously solve for the side-condition $\int_{x_0}^{x_1} G(y(x), y'(x), x)\, dx = 0$. If this approach is

possible, the Lagrange Lemma guarantees that the resulting y^* fulfills the isoperimetric side-condition. This approach is demonstrated using a classic example below.

It should be noted that the functions maximizing the functional of interest of the VML and VB frameworks are usually defined on the entire real line, rather than on a real interval with fixed endpoint conditions, as considered here. For simplicity, however, a formal treatment of the ensuing improper integrals is eschewed, and the methods developed in the current section are transferred to the problems considered in the next section by the intuition that the probability density functions of interest are usually Gaussian for which $f(x) \to 0$ in the limits of $x \to \infty$ and $x \to -\infty$.

The Lagrange Lemma for the isoperimetric variational problem

If the existence of a minimum $\hat{y}$ and of the corresponding Lagrangian multipliers λ_i $(i = 1, \dots, m)$ and μ_j $(j = 1, \dots, .p)$ is postulated, the following Lemma can readily be proven, and provides an ansatz for the solution of the isoperimetic variational problem:

Lagrange-Lemma for the isoperimetric variational problem
Let M be an arbitrary set and $J: M \to \mathbb{R}$ a function to be minimized on the restriction set

$$R := \{y \in M | g(y) = 0, h(y) \leq 0\}$$

where $g := (g_1, \dots, g_m): M \to \mathbb{R}^m$ and $h := (h_1, \dots, h_p): M \to \mathbb{R}^p$ define a set of side conditions. Further, let $\lambda_i \in \mathbb{R}$ $(i = 1, \dots, m)$ and $\mu_j \geq 0$ $(j = 1, \dots, .p)$ be specified such, that

$$\hat{y} \in \bar{S} := \left\{ y \in S \middle| \sum_{j=1}^{p} \mu_j h_j(y) = 0 \right\}$$

is a minimum of

$$\bar{\mathcal{F}} := J + \sum_{i=1}^{m} \lambda_i g_i + \sum_{j=1}^{p} \mu_j h_j : M \to \mathbb{R}$$

Then $\hat{y}$ is a Minimum of J on R.

The Lagrange Lemma for the isoperimetric variational problem can be verified by considering the following:

$$\begin{aligned} J(\hat{y}) &= J(\hat{y}) + 0 + 0 \\ &= J(\hat{y}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{y}) + \sum_{j=1}^{p} \mu_j h_j(\hat{y}) \\ &\leq J(y) + \sum_{i=1}^{m} \lambda_i g_i(y) + \sum_{j=1}^{p} \mu_j h_j(y) \\ &\leq J(y) \end{aligned} \tag{4.44}$$

If, like in the isoperimetric problem, the side conditions are given by equalities, the functions $h_j: M \to \mathbb{R}$ may be set to zero in the Lemma above, the restriction set simplifies to

$$R := \{y \in M | g(y) = 0\} \tag{4.45}$$

and the set $\bar{S}$ may be discarded. Thus, if the problem of minimizing the extended functional

$$\tilde{\mathcal{F}} := \int_{x_0}^{x_1} F(y(x), y'(x), x)\, dx \tag{4.46}$$

where

$$F(y(x), y'(x), x) := L(y(x), y'(x), x) + \lambda G(y(x), y'(x), x) \tag{4.47}$$

as specified in equations (4.46) and (4.47) can be solved, its solution provides a minimum of $\mathcal{F}$ which fulfills the side condition specified by I.

Example: The Problem of Dido

The problem of Dido may be stated as an isoperimetric variational problem as follows. Consider the vector space

$$V \coloneqq \{y \in \mathcal{C}^1[a,b] | y(a) = 0 = y(b)\} \tag{4.48}$$

The aim is to find a $y \in V$, which minimizes the functional

$$\mathcal{F}: V \to \mathbb{R}, y \mapsto \mathcal{F}(y) \coloneqq -\int_a^b y(x)\, dx \tag{4.49}$$

subject to the constraint (side-condition) that

$$g(y) = 0, \text{ where } g: V \to \mathbb{R}, y \mapsto g(y) \coloneqq \int_a^b \sqrt{1 + y'(x)^2}\, dx - l \tag{4.50}$$

for some constant $l \in \mathbb{R}$. Intuitively, this problem may be stated as maximizing the area between the graph of y and the x-axis using a y of a given constant arc length l (Figure 2, panels C and D). Applying the recipe for the solution of the isoperimetric variational problem discussed above, one obtains the following extended functional (see below for details):

$$\tilde{\mathcal{F}} \coloneqq -\int_a^b y(x)\, dx + \lambda \int_a^b \sqrt{1 + y'(x)^2}\, dx - l = \int_a^b \left(-y(x) + \lambda\sqrt{1 + y'(x)^2} - \frac{\lambda l}{b-a}\right) dx \tag{4.51}$$

where the Lagrange function is not explicitly dependent on x and is given by

$$L: \mathbb{R} \times \mathbb{R},\ (y, y') \mapsto L(y, y') = -y + \lambda\sqrt{1 + y'^2} - \frac{\lambda l}{b-a} \tag{4.52}$$

For a Lagrange function that explicitly only depends on y and y', the Euler-Lagrange equation simplifies to (see below for details)

$$L(y, y') - \frac{\partial}{\partial y'} L(y, y') y' = B \tag{4.53}$$

where $B \in \mathbb{R}$ is a constant. Evaluation of the Euler-Lagrange equation for the problem of Dido results in the following first-order ordinary differential equation for y (see below for details):

$$\sqrt{1 + y'^2}\left(-y - \frac{\lambda l}{b-a} - B\right) + \lambda = 0 \tag{4.54}$$

A solution for this differential equation is given by

$$y(x) = C + \sqrt{\lambda^2 - (x - D)^2} \tag{4.55}$$

where $C \coloneqq -B - \frac{\lambda l}{b-a}$ and $D \in \mathbb{R}$ is an additional integration constant (for a derivation of this solution, see (Richter, Mathias, o. J.)). Taking the square of the equation above then results in

$$(y(x) - C)^2 + (x - D)^2 = \lambda^2 \tag{4.56}$$

which is recognized as the functional equation for a circle with the center at $(D, C) \in \mathbb{R}^2$ and radius λ, while further analytical considerations allow for inferring that $D = \frac{a+b}{2}$ and $C \geq 0$ (Richter, Mathias, o. J.). The solution to the problem of Dido is thus a circular arc, the exact form of which depends on the relationship between the endpoints a, b and the given arc length l of y (Figure 2, panels C and D).

○ **Derivation of equation (4.53)**

To show that indeed

$$L(y,y',x) - \frac{\partial}{\partial y'}L(y,y',x)y' = B \tag{4.57}$$

for a real constant $B \in \mathbb{R}$, we show that the total deriviative of the function

$$f(y,y',x) \coloneqq L(y,y',x) - y'\frac{\partial}{\partial y'}L(y,y',x) \tag{4.58}$$

vanishes, provided that L is not an explicit function of x, and thus $\frac{\partial}{\partial x}L = \frac{\partial^2}{\partial y'\partial x}L = 0$:

$$\begin{aligned}\frac{d}{dx}f &= \frac{d}{dx}\left(L - y'\frac{\partial}{\partial y'}L\right) = \frac{d}{dx}L - \frac{d}{dx}\left(y'\frac{\partial}{\partial y'}L\right) \\ &= \left(\frac{\partial}{\partial x}L + y'\frac{\partial}{\partial y}L + y''\frac{\partial}{\partial y'}L\right) - \left(y''\frac{\partial}{\partial y'}L + y'\frac{\partial^2}{\partial x\partial y'}L + y'^2\frac{\partial^2}{\partial y\partial y'}L + y'y''\frac{\partial^2}{\partial y'^2}L\right) \\ &= y'\left(\frac{\partial}{\partial y}L - \frac{d}{dx}\frac{\partial}{\partial y'}L\right) \\ &= 0\end{aligned} \tag{4.59}$$

where the last equation follows from equation (4.54).

○ **Derivation of equation (4.54)**

We apply

$$L(y,y') - \frac{\partial}{\partial y'}L(y,y')y' = B \tag{4.60}$$

to the Lagrange function given in (4.52)

$$L(y,y') \coloneqq -y + \lambda\sqrt{1+y'^2} - \frac{\lambda l}{b-a} \tag{4.61}$$

and obtain

$$\begin{aligned}&-y + \lambda\sqrt{1+y'^2} - \frac{\lambda l}{b-a} - y'\frac{\partial}{\partial y'}\left(-y + \lambda\sqrt{1+y'^2} - \frac{\lambda l}{b-a}\right) = B \\ &\Leftrightarrow -y + \lambda\sqrt{1+y'^2} - \frac{\lambda l}{b-a} - y'\frac{\lambda y\prime}{\sqrt{1+y'^2}} = B\end{aligned} \tag{4.62}$$

Multiplying both sides by $\sqrt{1+y'^2}$ yields

$$\begin{aligned}&-y\sqrt{1+y'^2} + \lambda(1+y'^2) - \frac{\lambda l}{b-a}\sqrt{1+y'^2} - \lambda y'^2 = B\sqrt{1+y'^2} \\ \Leftrightarrow\ &-y\sqrt{1+y'^2} + \lambda + \lambda y'^2 - \frac{\lambda l}{b-a}\sqrt{1+y'^2} - \lambda y'^2 - B\sqrt{1+y'^2} = 0 \\ &\Leftrightarrow \sqrt{1+y'^2}\left(-y - \frac{\lambda l}{b-a} - B\right) + \lambda = 0\end{aligned} \tag{4.63}$$

□

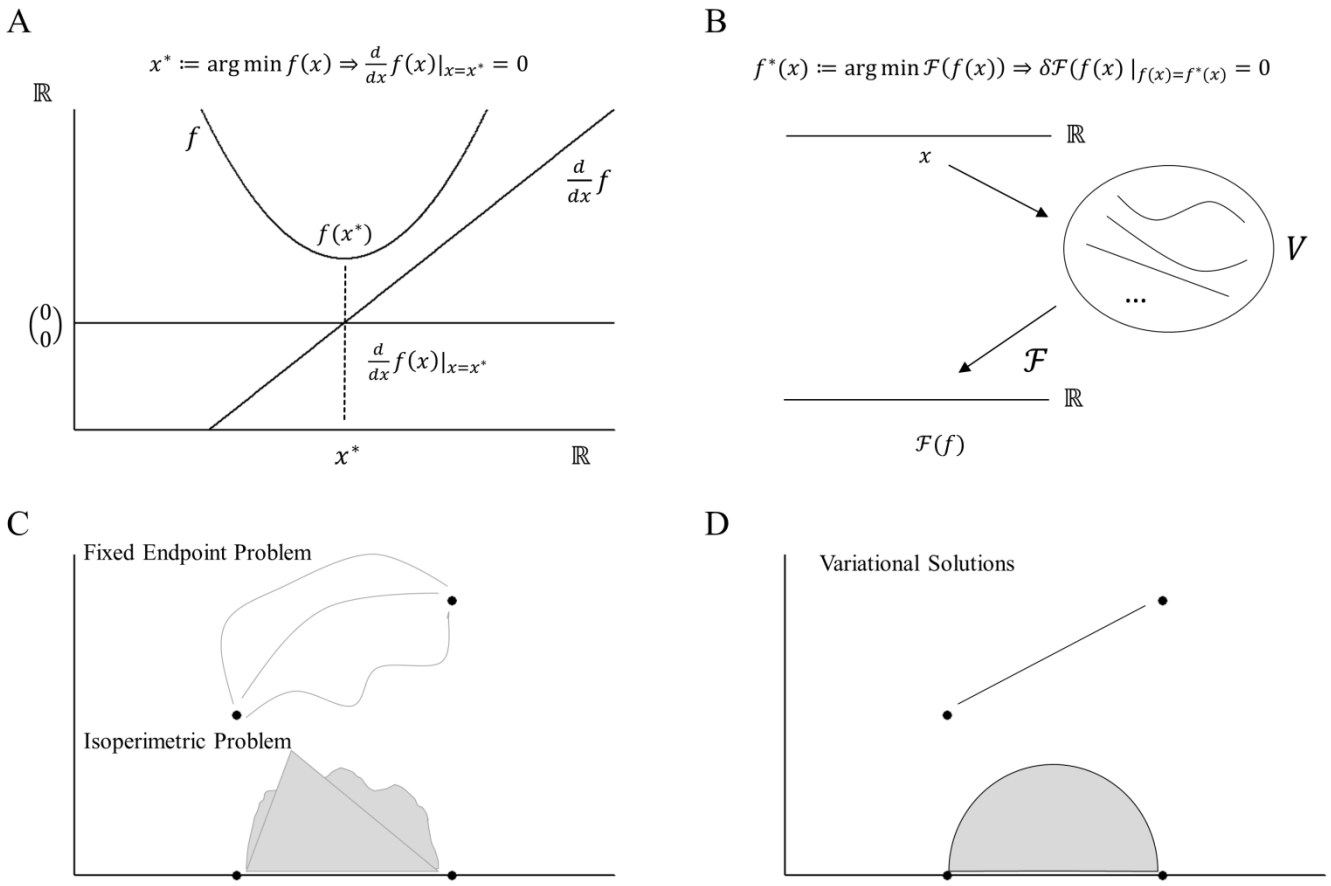

Figure 2. An intuitive notion of the calculus of variations. Figure 2A depicts a standard calculus analogy to the extremal problem of variational calculus. To find the extremal points x^* from a given real-valued function f, standard calculus usually proceeds by finding its stationary points. The stationary points of a function are those arguments of the function for which the derivative $\frac{d}{dx} f$ of the function vanishes, i.e., for which $\frac{d}{dx} f(x)|_{x=x^*} = 0$ holds. Figure 2B depicts a conceptual overview of the variational calculus framework: central to variational calculus is the problem of identifying a (real-valued) function f in a function space V, which extremizes a given functional $\mathcal{F}$. The functionals considered in this review (especially the variational free energy) map functions (in the context of the tutorial, probability density functions) onto the real line. Candidate functions are found by capitalizing on the necessary condition for an extremal, i.e., a vanishing Gateaux derivative $\delta\mathcal{F}(f(x))|_{f(x)=f^*(x)} = 0$ and solving for $f^*(x)$. Figures 2C and 2D sketch the variational problems and the solutions to them considered as examples in Section 3. The fixed endpoint problem can be conceived as the problem to find the curve (i.e., a mapping $f: \mathbb{R} \to \mathbb{R}^n$, here, $n = 2$) of minimal length connecting to given points in $\mathbb{R}^2$. The application of the variational calculus method allows for identifying the straight line connecting both points as the optimal function (curve). The isoperimetric problem (here, with fixed endpoints) represents a variational problem with additional integral side-conditions. The problem of Dido lies in finding the function of given arc length that maximizes the area between the function and the ordinate. Using the calculus of variations and a Lagrange multiplier approach, it can be shown that the solutions are circular arcs whose exact form depends on the predetermined arc length. Maximization of the variational free energy in the VB framework may be viewed as an isoperimetric problem. Here, the functions maximizing the variational free energy need to fulfill the integral side-condition of being probability density functions on the real line, i.e., their integral over the real line should equal 1.

Derivation of the VB inference theorem for mean-field approximation

In the VB framework the aim is to approximate the log marginal likelihood $\ln p(y)$ by iteratively maximizing its lower bound $\mathcal{F}\left(q(\vartheta_s), q\left(\vartheta_{\setminus s}\right)\right)$ with respect to the arguments $q(\vartheta_s)$ and $q\left(\vartheta_{\setminus s}\right)$. For iterations $i = 1,2, \ldots$, during the first maximization of finding $q^{(i+1)}(\vartheta_s)$, $q^{(i)}\left(\vartheta_{\setminus s}\right)$ is treated as a constant, while during the second maximization of finding $q^{(i+1)}(\vartheta_s)$, $q^{(i+1)}\left(\vartheta_{\setminus s}\right)$ is treated as a constant. However, as ϑ_s and $\vartheta_{\setminus s}$ may be used interchangeably, we here concern ourselves only with the case of maximizing $\mathcal{F}\left(q(\vartheta_s), q\left(\vartheta_{\setminus s}\right)\right)$ with Respect To $q(\vartheta_s)$. To obtain an expression for $q^{(i+1)}(\vartheta_s)$, we thus consider the functional

$$\mathcal{F}\left(q(\vartheta_s), q^{(i)}\left(\vartheta_{\setminus s}\right)\right) = \iint q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln\left(\frac{p(y,\vartheta)}{q(\vartheta) q^{(i)}\left(\vartheta_{\setminus s}\right)}\right) d\vartheta d\vartheta_{\setminus s}, \text{ where } \int q(\vartheta) d\vartheta = 1 \tag{4.64}$$

The extended Lagrange function is given in this case as

$$F\left(q(\vartheta_s), q^{(i)}\left(\vartheta_{\setminus s}\right)\right) = \int q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln\left(\frac{p(,y,\vartheta)}{q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right)}\right) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda \tag{4.65}$$

and, as in the previous section, the Gateaux derivative $\delta\mathcal{F}\left(q(x), q^{(i)}(\theta)\right)$ is given by the derivative of F with respect to $q(\vartheta_s)$, as $\mathcal{F}$ is not a function of $q'(\vartheta_s)$. One thus obtains with a constant $C \in \mathbb{R}$ (see below for details)

$$\delta\mathcal{F}\left(q(\vartheta_s), q^{(i)}\left(\vartheta_{\setminus s}\right)\right) = \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} - \ln q(\vartheta_s) + C \tag{4.66}$$

○ **Derivation of equation (4.66)**

$$\delta\mathcal{F}\left(q(\vartheta_s), q^{(i)}\left(\vartheta_{\setminus s}\right)\right) \tag{4.67}$$

$$= \frac{\partial}{\partial q(\vartheta_s)}\left(\int q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln\left(\frac{p(y,\vartheta)}{q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right)}\right) d\theta + \lambda q(\vartheta_s) - \lambda\right)$$

$$= \frac{\partial}{\partial q(\vartheta_s)}\left(\int q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} - \int q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln\left(q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right)\right) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda\right)$$

$$= \frac{\partial}{\partial q(\vartheta_s)}\left(q(\vartheta_s) \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} - \int q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln q(\vartheta_s)\, d\vartheta_{\setminus s} - \int q(\vartheta_s) q^{(i)}\left(\vartheta_{\setminus s}\right) \ln q\left(\vartheta_{\setminus s}\right) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda\right)$$

$$= \frac{\partial}{\partial q(\vartheta_s)}\left(q(\vartheta_s) \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} - q(\vartheta_s) \ln q(\vartheta_s) \int q^{(i)}\left(\vartheta_{\setminus s}\right) d\vartheta_{\setminus s} - q(\vartheta_s) \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln q\left(\vartheta_{\setminus s}\right) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda\right)$$

$$= \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} - \ln q(\vartheta_s) - 1 - \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln q^{(i)}\left(\vartheta_{\setminus s}\right) d\vartheta_{\setminus s} + \lambda$$

$$= \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} - \ln q(\vartheta_s) + C$$

□

Setting the Gateaux derivative (4.63) to zero thus yields

$$\ln q^{(i+1)}(\vartheta_s) \coloneqq \int q^{(i)}\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s} + C \tag{4.68}$$

Taking the exponential, expressing the multiplicative constant as proportionality factor then yields "VB inference theorem for mean-field approximations"

$$q^{(i+1)}(\vartheta_s) \propto \exp\left(\int q\left(\vartheta_{\setminus s}\right) \ln p(y,\vartheta)\, d\vartheta_{\setminus s}\right) \tag{4.69}$$

5 The Variational Maximum Likelihood Framework and its Application to LGSSMs

----

---------- Currently under extensive revision ----------

----

6 Mathematical details of the univariate Gaussian example

6.1 Introduction

To demonstrate the VB inference method for i.i.d. data we consider the Bayesian estimation of the expectation and variance parameters of a univariate Gaussian (discussed in detail in (Michael Chappell, Adrian Groves and Mark Woolrich, 2008; W. D. Penny, 2000) and visualized in Figure 5 of themain tutorial). We thus assume that the data is generated by a univariate Gaussian distribution over an observed variable y with true, but unknown, expectation parameter μ and inverse variance parameter λ, i.e.,

$$p(y) \coloneqq N(y; \mu, \lambda^{-1}) \tag{6.1}$$

Further, we assume that a set of N i.i.d. realizations $y_1^*, \dots, y_N^*$ of the observed variable y has been obtained. In this scenario, a classical maximum likelihood point estimation approach would result in estimates for μ and λ^{-1} based on the sample mean $\bar{\mu}$ and the (biased) sample variance $\bar{\sigma}^2$

$$\bar{\mu} = \sum_{n=1}^{N} y_n \text{ and } \bar{\sigma}^2 = \frac{1}{n}\sum_{n=1}^{N}(y_n - \bar{\mu})^2 \tag{6.2}$$

respectively. As outlined in the tutorial, based on appropriately chosen prior distributions, the aim of the Bayesian paradigm is to obtain posterior probability distributions, which quantify the remaining uncertainty over the true, but unknown, unobserved variables given the observed variables and an approximation to the model evidence, i.e. the probability of the data given the generative model. The generative model of the current example thus constitutes a joint probability distribution over the observed variable y and both unobserved random variables μ and $\lambda, \lambda > 0$. It is thus given by

$$p(y, \mu, \lambda) = p(\mu, \lambda)p(y|\mu, \lambda) = p(\mu, \lambda)\prod_{n=1}^{N} N\left(y_n|\mu, \lambda^{-1}\right) \tag{6.3}$$

A possible choice for a prior joint distribution over the unobserved variables is given by the product of a univariate Gaussian distribution over μ and a Gamma distribution over λ according to

$$(y, \mu, \lambda) = p(\mu)p(\lambda)p(y|\mu, \lambda) = N(\mu; m_\mu, s_\mu^2)G(\lambda; a_\lambda, b_\lambda)\prod_{n=1}^{N} N\left(y_n|\mu, \lambda^{-1}\right) \tag{6.4}$$

We consider a mean-field approximation to the posterior distribution over the data conditional unobserved variables, i.e., we set

$$p(\mu, \lambda|y) \approx q(\mu)q(\lambda) \tag{6.5}$$

Specifically, we set $q(\mu)$ to a normal distribution overμ with variational parameters m_μ^q and $s_\mu^{2^q}$, and we set $q(\lambda)$ to a Gamma distribution over λ with variational parameters a_λ^q, b_λ^q resulting in

$$q(\mu)q(\lambda) = N\left(x; m_\mu^q, s_\mu^{2^q}\right)G\left(\lambda; a_\lambda^q, b_\lambda^q\right) \tag{6.6}$$

In (6.6), $\left\{m_\mu^q, s_\mu^{2^q}, a_\lambda^q, b_\lambda^q\right\}$ represents a set of "variational" parameters. For convenience and to keep the notational overhead at bay, the values of the variational parameters on the ith iteration of the VB algorithm derived below will be denoted by $m_\mu^{(i)}, s_\mu^{2^{(i)}}, a_\lambda^{(i)}$, and $b_\lambda^{(i)}$, dropping the "q" superscript. To summarize, we have introduced a set of prior parameters $\left\{m_\mu, s_\mu^2, a_\lambda, b_\lambda\right\}$ governing the prior distribution $p(\mu, \lambda)$ of the generative model in (6.3) and a set

of of variational parameters $\left\{m_\mu^{(i)}, s_\mu^{2^{(i)}}, a_\lambda^{(i)}, b_\lambda^{(i)}\right\}$ governing the variational approximation $q(\mu)q(\lambda)$ to the posterior distribution $p(\mu, \lambda|y)$.

6.2 Application of the variational inference for approximate posteriors

As discussed in the tutorial the variational inference theorem for mean-field approximations (cf. equation (17)) states that the variational distribution over the unobserved variable partition ϑ_s is given by

$$q(\vartheta_s) \propto \exp\left(\int q\left(\vartheta_{\setminus s}\right)\ln p(y, \vartheta)\, d\vartheta_{\setminus s}\right) \tag{6.7}$$

For the current example, we have the mean field approximation

$$q(\mu, \lambda) = q(\mu)q(\lambda) \tag{6.8}$$

and thus

$$q(\mu) = C_\mu \cdot \exp(\int q(\lambda) \ln p(y, \mu, \lambda)\, d\lambda) \tag{6.9}$$

and

$$q(\lambda) = C_\lambda \cdot \exp(\int q(\mu) \ln p(y, \mu, \lambda)\, d\mu) \tag{6.10}$$

where C_μ and C_λ denote proportionality constants that render the proportionality statement in (6.7) equations in (6.9) and (6.10). In the following, we derive an iterative scheme based on the equations above. For this purpose, it is helpful (1.) to denote expectations by the expectation operator

$$\langle f(x)\rangle_{p(x)} \coloneqq \int p(x)f(x)dx \tag{6.11}$$

because it considerably reduces the visual complexity of the expressions involved, (2.) explicitly denote the iterative nature of the approach by indexing the variational distributions $q(\mu)$ and $q(\lambda)$ as $q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$, which also stresses the fact that in (6.9) and (6.10) above, the left hand side variational distribution refers to the $(i+1)$th iteration, while the right hand side variational distribution refers to the ith iteration, and (3.), because we are dealing with probability density functions of the exponential family, to log transform the expressions above to simplify proceedings. Upon these notational changes, (6.9) and (6.10) can be re-expressed as, in expectation-maximization algorithm terms as

E Step

$$\ln q^{(i+1)}(\mu) = \langle p(y, \mu, \lambda)\rangle_{q^{(i)}(\lambda)} + C_\mu \tag{6.12}$$

M Step

$$\ln q^{(i+1)}(\lambda) \coloneqq \langle p(y, \mu, \lambda)\rangle_{q^{(i+1)}(\mu)} + C \tag{6.13}$$

In the following two sections, we will derive explicit expressions for the variational parameters $m_\mu^{(i+1)}, s_\mu^{2^{(i+1)}}$ and $a_\lambda^{(i+1)}, b_\lambda^{(i+1)}$ that govern the variational distributions $q^{(i+1)}(\mu)$ and $q^{(i+1)}(\lambda)$, respectively. These expressions are given in terms of the observed data, as well as the preceeding variational parameters $m_\mu^{(i)}, s_\mu^{2^{(i)}}, a_\lambda^{(i)}$ and $b_\lambda^{(i)}$ as will be seen below.

6.3 Variational parameter update-equation derivation for the E Step

As noted above, for the E-Step, the variational inference theorem for mean-field approximations states that the logarithm of the optimal distribution over the unobserved variable μ, $\ln q^{(i+1)}(\mu)$, given a fixed distribution over the parameter variable, $q^{(i)}(\lambda)$, is given by

$$\ln q^{(i+1)}(\mu) = \langle p(y, \mu, \lambda) \rangle_{q^{(i)}(\lambda)} + C_\mu \tag{6.14}$$

which, if we identify the prior distribution $p(\mu)p(\lambda)$ with the initial variational distribution $q^{(0)}(\mu)q^{(0)}(\lambda)$ may be rewritten as

$$\ln q^{(i+1)}(\mu) = -\frac{1}{2}\langle \lambda \rangle_{q^{(i)}(\lambda)} \sum_{n=1}^{N} (y_n - \mu)^2 - \frac{\left(\mu - m_\mu^{(i)}\right)^2}{2s_\mu^{2\,(i)}} + C_\mu \tag{6.15}$$

Based on (6.15) and the completing-the-square theorem (Supplement Section 2), one can infer that the optimal variational distribution over the latent variable μ is proportional to a Gaussian

$$q^{(i+1)}(\mu) \propto N\left(\mu; m_\mu^{(i+1)}, s_\mu^{2\,(i+1)}\right) \tag{6.16}$$

where the updated variational parameters are provided in terms of the data $y_1, \ldots, y_N$ and the variational parameters of $q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$ as

$$m_\mu^{(i+1)} \coloneqq \frac{m_\mu^{(i)} + s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^{N} y_n}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}} \quad \text{and} \quad s_\mu^{2\,(i+1)} \coloneqq \frac{s_\mu^{2\,(i)}}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}} \tag{6.17}$$

- **Derivation of equation (6.15)**

For a fixed distribution $q^{(i)}(\lambda)$, we have due to the i.i.d. data assumption and the assumption of factorized priors for $y = y_{1:N}$

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= \langle p(y, \mu, \lambda) \rangle_{q^{(i)}(\lambda)} + C_\mu \\ &= \langle \ln(\textstyle\prod_{n=1}^{N} p(y_n, \mu, \lambda)) \rangle_{q^{(i)}(\lambda)} + C_\mu \\ &= \langle \ln(\textstyle\prod_{n=1}^{N} p(y_n|\mu, \lambda)\, p(\mu, \lambda)) \rangle_{q^{(i)}(\lambda)} + C_\mu \\ &= \langle \ln(\textstyle\prod_{n=1}^{N} p(y_n|\mu, \lambda)\, p(\mu)p(\lambda)) \rangle_{q^{(i)}(\lambda)} + C_\mu \end{aligned} \tag{6.18}$$

Setting

$$p(\mu)p(\lambda) = q^{(i)}(\mu)q^{(i)}(\lambda) \text{ for } i = 0 \tag{6.19}$$

i.e. initializing the variational distribution with the prior distribution, we have for iterationsr $i = 0, 1, 2, \ldots$.

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= \langle \ln\big(\textstyle\prod_{n=1}^{N} p(y_n|\mu, \lambda)\, q^{(i)}(\mu)q^{(i)}(\lambda)\big) \rangle_{q^{(i)}(\lambda)} + C_\mu \\ &= \langle \ln(\textstyle\prod_{n=1}^{N} p(y_n|\mu, \lambda)) \rangle_{q^{(i)}(\lambda)} + \langle \ln q^{(i)}(\mu) \rangle_{q^{(i)}(\lambda)} + \langle \ln q^{(i)}(\lambda) \rangle_{q^{(i)}(\lambda)} + C_\mu \end{aligned} \tag{6.20}$$

Substituting the example specific functional forms of $p(y_n|\mu,\lambda), q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$ from (6.4) and (6.6) above, namely

$$p(y_n|\mu,\lambda) \coloneqq \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(y_n-\mu)^2\right) \tag{6.21}$$

$$q^{(i)}(\mu) \coloneqq \left(2\pi s_\mu^{2\,(i)}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_\mu^{2\,(i)}}\left(\mu - m_\mu^{(i)}\right)^2\right) \tag{6.22}$$

and

$$q^{(i)}(\lambda) \coloneqq \frac{1}{\Gamma\left(a_\lambda^{(i)}\right)} \frac{1}{\left(b_\lambda^{(i)}\right)^{a_\lambda^{(i)}}} \lambda^{a_\lambda^{(i)}-1} \exp\left(-\frac{\lambda}{b_\lambda^{(i)}}\right) \tag{6.23}$$

in (6.20) then yields

$$\begin{aligned}
\ln q^{(i+1)}(x) &= \langle \ln\left(\prod_{n=1}^N \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(y_n-\mu)^2\right)\right)\rangle_{q^{(i)}(\lambda)} \\
&\quad + \langle \ln\left(\left(2\pi s_\mu^{2\,(i)}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_\mu^{2\,(i)}}\left(\mu - m_\mu^{(i)}\right)^2\right)\right)\rangle_{q^{(i)}(\lambda)} \\
&\quad + \langle \ln\left(\frac{1}{\Gamma\left(a_\lambda^{(i)}\right)} \frac{1}{\left(b_\lambda^{(i)}\right)^{a_\lambda^{(i)}}} \lambda^{a_\lambda^{(i)}-1} \exp\left(-\frac{\lambda}{b_\lambda^{(i)}}\right)\right)\rangle_{q^{(i)}(\lambda)} + C_\mu \\
&= \langle \frac{N}{2}\ln\lambda - \frac{1}{2}\ln 2\pi - \frac{\lambda}{2}\sum_{n=1}^N (y_n - x)^2\rangle_{q^{(i)}(\lambda)} \\
&\quad + \langle -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln s_\mu^{2\,(i)} - \frac{\left(\mu - m_\mu^{(i)}\right)^2 1}{2s_\mu^{2\,(i)}}\rangle_{q^{(i)}(\lambda)} \\
&\quad + \langle -\ln\left(\Gamma\left(a_\lambda^{(i)}\right)\right) - \ln(b_\lambda^{(i)})^{a_\lambda^{(i)}} + \left(a_\lambda^{(i)} - 1\right)\ln\lambda - \frac{\lambda}{b_\lambda^{(i)}})\rangle_{q^{(i)}(\lambda)} + C_\mu
\end{aligned} \tag{6.24}$$

Grouping all constant terms independent of μ (i.e. those terms that do not change if μ changes) with the constant C_μ and evaluation of the expectation then simplifies the above to

$$\begin{aligned}
\ln q^{(i+1)}(\mu) &= -\langle \frac{\lambda}{2}\sum_{n=1}^N (y_n-\mu)^2\rangle_{q^{(i)}(\lambda)} - \langle \frac{\left(\mu - m_\mu^{(i)}\right)^2}{2s_\mu^{2\,(i)}}\rangle_{q^{(i)}(\lambda)} + C_\mu \\
&= -\frac{1}{2}\langle\lambda\rangle_{q^{(i)}(\lambda)} \sum_{n=1}^N (y_n-\mu)^2 - \frac{\left(\mu - m_\mu^{(i)}\right)^2}{2s_\mu^{2\,(i)}} + C_\mu
\end{aligned} \tag{6.25}$$

□

- **Derivation of equations (6.16) and (6.17)**

We first note that the expectation of λ under $q^{(i)}(\lambda)$ may be expressed in terms of the variational parameters as

$$\langle\lambda\rangle_{q^{(i)}(\lambda)} = a_\lambda^{(i)} b_\lambda^{(i)} \tag{6.26}$$

and the term $\sum_{n=1}^N (y_n-\mu)^2$ may be rewritten as

$$\sum_{n=1}^N (y_n-\mu)^2 = \sum_{n=1}^N (y_n^2 - 2y_n\mu + \mu^2) = \sum_{n=1}^N y_n^2 - 2\mu\sum_{n=1}^N y_n + N\mu^2 \tag{6.27}$$

By reordering (6.15) in terms of powers of μ we thus obtain

$$\ln q^{(i+1)}(\mu) = -a_\lambda^{(i)} b_\lambda^{(i)} \left(\sum_{n=1}^N y_n^2 - 2\mu \sum_{n=1}^N y_n + N\mu^2\right) - \frac{\mu^2 - 2\mu m_\mu^{(i)} + \left(m_\mu^{(i)}\right)^2}{2s_\mu^{2\,(i)}} + C_\mu \tag{6.28}$$

$$= -a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n^2 + \mu a_\lambda^{(i)} b_\lambda^{(i)} 2 \sum_{n=1}^N y_n - a_\lambda^{(i)} b_\lambda^{(i)} N\mu^2 - \frac{\mu^2}{2s_\mu^{2\,(i)}} - \frac{\mu m_\mu^{(i)}}{s_\mu^{2\,(i)}} + \frac{\left(m_\mu^{(i)}\right)^2}{2s_\mu^{2\,(i)}} + C_\mu$$

$$= -N a_\lambda^{(i)} b_\lambda^{(i)} \mu^2 - \frac{1}{2s_\mu^{2\,(i)}} \mu^2 + \left(2a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n\right) \mu - \frac{m_\mu^{(i)}}{s_\mu^{2\,(i)}} \mu + C_\mu$$

$$= -\frac{1}{2}\left(N a_\lambda^{(i)} b_\lambda^{(i)} + \frac{1}{s_\mu^{2\,(i)}}\right) \mu^2 + \left(2a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n - \frac{m_\mu^{(i)}}{s_\mu^{2\,(i)}}\right) \mu + C_\mu$$

$$= -\frac{1}{2}\left(\frac{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}}{s_\mu^{2\,(i)}}\right) \mu^2 + \left(2a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n - \frac{m_\mu^{(i)}}{s_\mu^{2\,(i)}}\right) \mu + C_\mu$$

Using the completing-the-square theorem in the form

$$\exp\left(-\frac{1}{2} a x^2 - b x\right) = N(x; a^{-1} b, a^{-1}) \cdot C \tag{6.29}$$

then yields

$$q^{(i+1)}(\mu) \propto N\left(\mu; m_\mu^{(i+1)}, s_\mu^{2\,(i+1)}\right) \tag{6.30}$$

where the parameters $m_\mu^{(i+1)}$ and $s_\mu^{2\,(i+1)}$ may be expressed in terms of the ith variational parameters as

$$s_\mu^{2\,(i+1)} \coloneqq \left(\frac{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}}{s_\mu^{2\,(i)}}\right)^{-1} = \frac{s_\mu^{2\,(i)}}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}} \tag{6.31}$$

and

$$m_\mu^{(i+1)} \coloneqq \frac{s_\mu^{2\,(i)}}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}} \left(a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n - \frac{m_\mu^{(i)}}{s_\mu^{2\,(i)}}\right) \tag{6.32}$$

$$= \frac{s_\mu^{2\,(i)}}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}} \left(\frac{m_\mu^{(i)} + a_\lambda^{(i)} b_\lambda^{(i)} s_\mu^{2\,(i)} \sum_{n=1}^N y_n}{s_\mu^{2\,(i)}}\right)$$

$$= \frac{m_\mu^{(i)} + s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}}$$

□

6.4 Variational parameter update-equation derivation for the E Step

For the M-Step, the variational inference theorem for approximated posteriors (6.13) states that the logarithm of the optimal distribution over the unobserved variable λ, given a momentarily fixed distribution over the unobserved variable μ variable, is obtained by setting (with a constant $C_\lambda \in \mathbb{R}$)

$$\ln q^{(i+1)}(\lambda) \coloneqq \langle p(y, \mu, \lambda) \rangle_{q^{(i+1)}(\mu)} + C_\lambda \tag{6.33}$$

which, if we identify the prior distribution $p(x)p(\lambda)$ with the initial variational distribution $q^{(0)}(x)q^{(0)}(\lambda)$, may be rewritten as

$$\ln q^{(i+1)}(\lambda) = \frac{N}{2} \ln \lambda - \frac{\lambda}{2} \langle \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i+1)}(\mu)} + \left(a_\lambda^{(i)} - 1\right) \ln \lambda - \frac{\lambda}{b_\lambda^{(i)}} + C_\lambda \tag{6.34}$$

Further, upon evaluation of the expectation and taking the exponential of (6.13), it follows that $q^{(i+1)}(\lambda)$ is proportional to a Gamma distribution

$$G\left(\lambda; a_\lambda^{(i+1)}, b_\lambda^{(i+1)}\right) \tag{6.35}$$

with parameters

$$a_\lambda^{(i+1)} \coloneqq \frac{N}{2} + a_\lambda^{(i)} \text{ and } b_\lambda^{(i+1)} \coloneqq \left(\frac{1}{b_\lambda^{(i)}} + \frac{1}{2}\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n\, \mu_x^{(i+1)} + N\left(\left(m_\mu^{(i+1)}\right)^2 + s_\mu^{2\,(i+1)}\right)\right)\right)^{-1} \tag{6.36}$$

- **Derivation of equation (6.34)**

In analogy to the derivation of equations (6.22) and (6.23) we have

$$\begin{aligned}\ln q^{(i+1)}(\lambda) = {} & \langle \ln\left(\prod_{n=1}^N \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(y_n-\mu)^2\right)\right)\rangle_{q^{(i+1)}(\mu)} \\ & + \langle \ln\left(\left(2\pi s_\mu^{2\,(i+1)}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_\mu^{2\,(i+1)}}\left(\mu - m_\mu^{(i+1)}\right)^2\right)\right)\rangle_{q^{(i+1)}(\mu)} \\ & + \langle \ln\left(\frac{1}{\Gamma\left(a_\lambda^{(i)}\right)} \frac{1}{(b_\lambda^{(i)})^{a_\lambda^{(i)}}} \lambda^{a_\lambda^{(i)}-1} \exp\left(-\frac{\lambda}{b_\lambda^{(i)}}\right)\right)\rangle_{q^{(i+1)}(x)} + C_\lambda \end{aligned} \tag{6.37}$$

$$\begin{aligned} = {} & \langle \frac{N}{2}\ln\lambda - \frac{1}{2}\ln 2\pi - \frac{\lambda}{2}\sum_{n=1}^N (y_n-\mu)^2\rangle_{q^{(i+1)}(\mu)} \\ & + \langle -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln s_\mu^{2\,(i+1)} - \frac{\left(\mu - m_\mu^{(i+1)}\right)^2}{2s_\mu^{2\,(i+1)}}\rangle_{q^{(i+1)}(\mu)} \\ & + \langle -\ln\left(\Gamma\left(a_\lambda^{(i)}\right)\right) - \ln\left(\left(b_\lambda^{(i)}\right)^{a_\lambda^{(i)}}\right) + \left(a_\lambda^{(i)}-1\right)\ln\lambda - \frac{\lambda}{b_\lambda^{(i)}}\rangle_{q^{(i+1)}(\mu)} + C_\lambda \end{aligned}$$

Grouping all constant terms independent of λ (i.e. those terms that do not change if λ changes) with the constant C_λ then simplifies the above to

$$\ln q^{(i+1)}(\lambda) = \langle \frac{N}{2}\ln\lambda - \frac{\lambda}{2}\sum_{n=1}^N (y_n-\mu)^2\rangle_{q^{(i+1)}(\mu)} + \langle\left(a_\lambda^{(i)}-1\right)\ln\lambda - \frac{\lambda}{b_\lambda^{(i)}}\rangle_{q^{(i+1)}(x)} + C_\lambda \tag{6.38}$$

Evaluation of the expectation of μ under $q^{(i+1)}(\mu)$ then yields

$$\ln q^{(i+1)}(\lambda) = \frac{N}{2}\ln\lambda - \frac{\lambda}{2}\langle\sum_{n=1}^N (y_n-x)^2\rangle_{q^{(i+1)}(\mu)} + \left(a_\lambda^{(i)}-1\right)\ln\lambda - \frac{\lambda}{b_\lambda^{(i)}} + C_\lambda \tag{6.39}$$

□

- **Derivation of equations (6.35) and (6.36)**

Summarizing the right hand side of equation (6.34) in terms of logarithms of λ yields

$$\begin{aligned}\ln q^{(i+1)}(\lambda) &= \left(\frac{N}{2} + a_\lambda^{(i)} - 1\right)\ln\lambda - \left(\frac{1}{b_\lambda^{(i)}} - \frac{1}{2}\langle\sum_{n=1}^N (y_n-\mu)^2\rangle_{q^{(i+1)}(\mu)}\right)\lambda + C_\lambda \\ &= \left(\frac{N}{2} + a_\lambda^{(i)} - 1\right)\ln\lambda - \left(\frac{1}{b_\lambda^{(i)}} - \frac{1}{2}\langle\sum_{n=1}^N y_n^2 - 2\mu\sum_{n=1}^N y_n + N\mu^2\rangle_{q^{(i+1)}(\mu)}\right)\lambda + C_\lambda \end{aligned} \tag{6.40}$$

$$= \left(\frac{N}{2} + a_\lambda^{(i)} - 1\right) \ln \lambda - \left(\frac{1}{b_\lambda^{(i)}} - \frac{1}{2}\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n \langle \mu \rangle_{q^{(i+1)}(\mu)} + N\langle \mu^2 \rangle_{q^{(i+1)}(\mu)}\right)\right)\lambda + C_\lambda$$

where the expectations may be expressed in terms of the variational parameters of $q^{(i+1)}(\mu)$ according to

$$\ln q^{(i+1)}(\lambda) = \left(\frac{N}{2} + a_\lambda^{(i)} - 1\right) \ln \lambda - \left(\frac{1}{b_\lambda^{(i)}} - \frac{1}{2}\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n m_\mu^{(i+1)} + N\left(\left(m_\mu^{(i+1)}\right)^2 + s_\mu^{2\,(i+1)}\right)\right)\right)\lambda + C_\lambda \tag{6.41}$$

Taking the exponential on both sides then yields

$$q^{(i+1)}(\lambda) \propto \lambda^{\left(\frac{N}{2} + a_\lambda^{(i)} - 1\right)} \exp\left(\left(-\frac{1}{b_\lambda^{(i)}} - \frac{1}{2}\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n m_\mu^{(i+1)} + N\left(\left(m_\mu^{(i+1)}\right)^2 + s_\mu^{2\,(i+1)}\right)\right)\right)\lambda\right) \tag{6.42}$$

Up to a normalization constant C_λ, $q^{(i+1)}(\lambda)$ is thus given by a Gamma distribution with

$$q^{(i+1)} \propto \lambda^{a_\lambda^{(i+1)}} \exp\left(-\frac{\lambda}{b_\lambda^{(i+1)}}\right) \tag{6.43}$$

where

$$a_\lambda^{(i+1)} \coloneqq \frac{N}{2} + a_\lambda^{(i)} \tag{6.44}$$

and

$$b_\lambda^{(i+1)} \coloneqq \left(\frac{1}{b_\lambda^{(i)}} + \frac{1}{2}\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n m_\mu^{(i+1)} + N\left(\left(m_\mu^{(i+1)}\right)^2 + s_\mu^{2\,(i+1)}\right)\right)\right)^{-1} \tag{6.45}$$

□

6.5 Evaluation of the variational free energy

While the maximization of the variational free energy has been exploited to derive the variational parameter update equations, the value of the variational free energy has not directly been evaluated. To obtain a value of this (implicit) target function to monitor the evolution of the algorithm, and, more importantly, to obtain an approximation to the marginal likelihood upon VB-EM convergence, it is desirable to evaluate the variational free energy integral (cf. equation (13) of the main tutorial). We first reformulate the variational free energy as follows (W. D. Penny, 2000)

$$\mathcal{F}\big(q(\vartheta)\big) = \int q(\vartheta) \ln\left(\frac{p(y|\vartheta)}{q(\vartheta)}\right) d\vartheta - \int q(\vartheta) \ln\left(\frac{q(\vartheta)}{p(\vartheta)}\right) d\vartheta \coloneqq \mathcal{L}_{av}\big(p(y|\vartheta), q(\vartheta)\big) - \mathcal{KL}\big(q(\vartheta)||p(\vartheta)\big) \tag{6.46}$$

The first term in the expression on the right hand side of (6.46) is often referred to as the "average energy" and the second term is the KL-divergence between the variational and prior distributions.

○ ***Derivation of (6.46)***

By definition of the variational free energy, we have

$$\mathcal{F}\big(q(\vartheta)\big) = \int q(\vartheta) \ln\left(\frac{p(y,\vartheta)}{q(\vartheta)}\right) d\vartheta \tag{6.47}$$

$$= \int q(\vartheta) \ln\left(\frac{p(y|\vartheta)p(\vartheta)}{q(\vartheta)}\right) d\vartheta$$

$$= \int q(\vartheta) \ln\left(\frac{p(y|\vartheta)}{q(\vartheta)} - \frac{q(\vartheta)}{p(\vartheta)}\right) d\vartheta$$

$$= \int q(\vartheta) \ln\left(\frac{p(y|\vartheta)}{q(\vartheta)}\right) d\vartheta - \int q(\vartheta) \ln\left(\frac{q(\vartheta)}{p(\vartheta)}\right) d\vartheta$$

$$= \mathcal{L}_{av}\big(p(y|\vartheta), q(\vartheta)\big) - \mathcal{KL}\big(q(\vartheta)||p(\vartheta)\big)$$

□

The "average energy term " $\mathcal{L}_{av}$ may further be reformulated as follows

$$\mathcal{L}_{av}\big(p(y|\vartheta), q(\vartheta)\big) = \int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta + \mathcal{H}(q(\vartheta)) \tag{6.48}$$

- ***Verification of (6.48)***

By definition of the average energy and entropy, we have

$$\mathcal{L}_{av} = \int q(\vartheta) \ln\left(\frac{p(y|\vartheta)}{q(\vartheta)}\right) d\vartheta \tag{6.49}$$

$$= \int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta - \int q(\vartheta) \ln q(\vartheta)\, d\vartheta$$

$$= \int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta + \mathcal{H}(q(\vartheta))$$

□

Upon this reformulation the respective terms comprising the variational free energy,

$$\mathcal{F}\big(q(\vartheta)\big) = \int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta + \mathcal{H}(q(\vartheta)) - \mathcal{KL}\big(q(\vartheta)||p(\vartheta)\big) \tag{6.50}$$

may then be evaluated in turn as shown below, resulting in

$$\mathcal{F}^{(i)}\left(y, N, m_\mu^{(i)}, s_\mu^{2\,(i)}, a_\lambda^{(i)}, b_\lambda^{(i)}\right) \coloneqq \mathcal{L}_{av} - KL(q^{(i)}(\mu)||p(\mu)) - KL(q^{(i)}(\lambda)||p(\lambda)) \tag{6.51}$$

where

$$\mathcal{L}_{av} \coloneqq \frac{1}{2}\left(\psi\left(a_\lambda^{(i)}\right) + \ln\left(b_\lambda^{(i)}\right)\right) - \frac{1}{2} a_\lambda^{(i)} b_\lambda^{(i)} \left(\sum_{n=1}^{N} y_n^2 + N\left(\left(m_\mu^{(i)}\right)^2 + s_\mu^{2\,(i)}\right) - 2m_\mu^{(i)} \sum_{n=1}^{N} y_n\right) \tag{6.52}$$

We thus obtainthus obtains an expression for the variational free energy as function of the observed data $y_1, \dots, y_n$, the variational parameters $m_\mu^{(i)}, s_\mu^{2\,(i)}, a_\lambda^{(i)}, b_\lambda^{(i)}$ and the prior variational parameter values $a_\lambda^{(0)}, b_\lambda^{(0)}$.

- ***Verification of (6.50)***

We first recall the definitions of $q(y|\vartheta), p(\vartheta)$ and $q(\vartheta)$ for the current example:

$$p(y|\vartheta) = p(y|\mu, \lambda) = \prod_{n=1}^{N} N\,(y_n|\mu, \lambda) \tag{6.53}$$

$$p(\vartheta) = p(\mu, \lambda) = p(x)p(\lambda) = N(\mu; m_\mu, s_\mu^2) G(\lambda; a_\lambda, b_\lambda) \tag{6.54}$$

and

$$q(\vartheta) = q(\mu, \lambda) = q(\mu)q(\lambda) = N\left(\mu; m_\mu^{(i)}, {s_\mu^2}^{(i)}\right) G(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}) \tag{6.55}$$

(a) Evaluation of $\int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta$ *in (6.50)*

Substitution of (6.53) in (6.50) yields

$$\iint q(\mu)q(\lambda) \ln\left(\prod_{n=1}^N N(y_n|\mu, \lambda)\right) d\mu d\lambda = \iint q(\mu)q(\lambda) \ln\left(\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \prod_{n=1}^N \exp\left(-\frac{\lambda}{2}(y_n - \mu)^2\right)\right) d\mu d\lambda \tag{6.56}$$

$$= \iint q(\mu)q(\lambda) \left(\frac{1}{2}\ln\lambda - \frac{1}{2}\ln 2\pi - \frac{\lambda}{2}\sum_{n=1}^N (y_n - \mu)^2\right)\, d\mu d\lambda$$

$$= \frac{1}{2}\iint q(\mu)q(\lambda)\, (\ln\lambda)\, d\mu d\lambda - \iint q(\mu)q(\lambda) \left(\frac{\lambda}{2}\sum_{n=1}^N (y_n - \mu)^2\right) d\mu d\lambda - \frac{1}{2}\ln 2\pi$$

$$= \frac{1}{2}\int q(\lambda) \ln\lambda\, d\lambda - \int q(\lambda)\frac{\lambda}{2}\left(\int q(\mu)\left(\sum_{n=1}^N (y_n - \mu)^2\right) d\mu\right) d\lambda - \frac{1}{2}\ln 2\pi$$

The first integral term above is the expectation of a logarithm variable λ under the variational Gamma distribution $G\left(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}\right)$ which can be shown to evaluate to

$$\int q(\lambda) \ln\lambda\, d\lambda = \psi\left(b_\lambda^{(i)}\right) + \ln a_\lambda^{(i)} \tag{6.57}$$

where

$$\psi: \mathbb{R} \to \mathbb{R}, x \mapsto \psi(x) \coloneqq \frac{d}{dx}\Gamma(x) \tag{6.58}$$

denotes the di-gamma function. The verification of (6.57) is left as an exercise to the interested reader. The second integral term in (6.56) evaluates to

$$\int q(\lambda)\frac{1}{2}\lambda\left(\int q(\mu)\left(\sum_{n=1}^N (y_n - \mu)^2\right) d\mu\right) d\lambda = \int q(\lambda)\left(\frac{1}{2}\lambda\right)\left(\int q(\mu)\left(\sum_{n=1}^N y_n^2 - 2\mu\sum_{n=1}^N y_n + N\mu^2\right) d\mu\right) d\lambda \tag{6.59}$$

$$= \frac{1}{2}\int q(\lambda)\, \lambda\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n \int q(\mu)\mu d\mu + N\int q(\mu)\mu^2 d\mu\right) d\lambda$$

$$= \frac{1}{2}a_\lambda^{(i)} b_\lambda^{(i)} \left(\sum_{n=1}^N y_n^2 - 2m_\mu^{(i)}\sum_{n=1}^N y_n + N\left(\left(\mu_x^{(i)}\right)^2 + {s_\mu^2}^{(i)}\right)\right)$$

(b) Evaluation of $\mathcal{H}(q(\vartheta))$ *in (6.50)*

It is a well-known result from information theory that the joint entropy of statistically independent random variables is additive (Cover, 1991). We thus have

$$\mathcal{H}\big(q(\vartheta)\big) = \mathcal{H}\big(q(\mu)q(\lambda)\big) = \mathcal{H}\big(q(\mu)\big) + \mathcal{H}\big(q(\lambda)\big) \tag{6.60}$$

The entropy of the variable μ corresponds to the differential entropy of the variational Gaussian distribution

$$q^{(i)}(\mu) = N\left(\mu; m_\mu^{(i)}, {s_\mu^2}^{(i)}\right) \tag{6.61}$$

and is thus given by (see e.g. Bishop)

$$\mathcal{H}\big(q(\mu)\big) = \mathcal{H}\left(N\left(\mu; m_\mu^{(i)}, {s_\mu^2}^{(i)}\right)\right) = \frac{1}{2}\ln {s_\mu^2}^{(i)} + \frac{1}{2}(1 + \ln 2\pi) \tag{6.62}$$

The entropy of the variable λ corresponds to the differential entropy of the variational Gamma distribution and is thus given by (see e.g. Bishop)

$$\mathcal{H}\big(q(\lambda)\big) = \mathcal{H}\left(G\left(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}\right)\right) = b_\lambda^{(i)} + \ln a_\lambda^{(i)} + \ln\left(\Gamma\left(b_\lambda^{(i)}\right)\right) + \left(1 - b_\lambda^{(i)}\right)\psi\left(b_\lambda^{(i)}\right) \tag{6.63}$$

(c) Evaluation of $\mathcal{KL}\big(q(\vartheta)||p(\vartheta)\big)$ in (6.50)

We first note that we have

$$\mathcal{KL}(q(\mu)q(\lambda)||p(\mu)p(\lambda)) = \mathcal{KL}(q(\mu)||p(\mu)) + \mathcal{KL}(q(\lambda)||p(\lambda)) \tag{6.64}$$

because

$$\mathcal{KL}(q(\mu)q(\lambda)||p(\mu)p(\lambda)) = \int q(\mu)q(\lambda)\ln\left(\frac{p(\mu)p(\lambda)}{q(\mu)q(\lambda)}\right)d\mu\, d\lambda \tag{6.65}$$

$$= \int q(\mu)q(\lambda)\left(\ln\frac{p(\mu)}{q(\mu)} + \ln\frac{p(\lambda)}{q(\lambda)}\right)d\mu\, d\lambda$$

$$= \int q(\mu)\left(\ln\frac{p(\mu)}{q(\mu)}\right)d\mu + \int q(\lambda)\left(\ln\frac{p(\lambda)}{q(\lambda)}\right)d\lambda$$

$$= \mathcal{KL}(q(\mu)||p(\mu)) + \mathcal{KL}(q(\lambda)||p(\lambda))$$

In the current example, the variable μ is governed by a Gaussian distribution. The prior distribution $p(\mu)$ corresponds to the initial variational $q^{(0)}(\mu)$, while the variational distribution $q(\mu)$ corresponds to the i variational distribution over μ. Hence, we have

$$\mathcal{KL}(q(\mu)||p(\mu)) = \mathcal{KL}\left(q^{(i)}(\mu)||q^{(0)}(\mu)\right) = \mathcal{KL}\left(N\left(\mu; m_\mu^{(i)}, {s_\mu^2}^{(i)}\right)||N\left(\mu; m_\mu^{(0)}, {s_\mu^2}^{(0)}\right)\right) \tag{6.66}$$

As noted in [Penny [xXXRef]], the KL divergence for two Gaussian distributions is a function of the respective distribution parameters and given for the current example as

$$\mathcal{KL}\left(N\left(\mu; m_\mu^{(i)}, {s_\mu^2}^{(i)}\right)||N\left(\mu; m_\mu^{(0)}, {s_\mu^2}^{(0)}\right)\right) = \frac{1}{2}\ln\frac{{s_\mu^2}^{(0)}}{{s_\mu^2}^{(i)}} + \frac{{m_\mu^{(0)}}^2 + {m_\mu^{(i)}}^2 + {s_\mu^2}^{(i)} - 2m_\mu^{(i)}m_\mu^{(0)}}{2{s_\mu^2}^{(0)}} \tag{6.67}$$

Finally, for the current example, the parameter variable corresponds to λ as is governed by a Gamma distribution. The prior distribution $p(\lambda)$ corresponds to the initial variational distribution $q^{(0)}(\lambda)$ while the posterior variational distribution corresponds to the ith variational distribution over λ.

$$\mathcal{KL}\big(q(\lambda)||p(\lambda)\big) = \mathcal{KL}\left(q^{(i)}(\lambda)||q^{(0)}(\lambda)\right) = \mathcal{KL}\left(G\left(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}\right)||G\left(\lambda; a_\lambda^{(0)}, b_\lambda^{(0)}\right)\right) \tag{6.68}$$

As noted in (Penny, 2001), the KL divergence for two Gamma distributions is a function of the respective Gamma distribution parameters and given for the current example as

$$\mathcal{KL}\left(G\left(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}\right)||G\left(\lambda; a_\lambda^{(0)}, b_\lambda^{(0)}\right)\right) \tag{6.69}$$

$$= \left(b_\lambda^{(i)} - 1\right)\psi\left(b_\lambda^{(i)}\right) - \ln a_\lambda^{(i)} - b_\lambda^{(i)} - \log\Gamma\left(b_\lambda^{(i)}\right) + \ln\Gamma\left(b_\lambda^{(0)}\right) + b_\lambda^{(0)} \ln a_\lambda^{(0)} - \left(b_\lambda^{(0)} - 1\right)\left(\psi\left(b_\lambda^{(i)}\right) + \ln a_\lambda^{(i)}\right) + \frac{a_\lambda^{(i)} b_\lambda^{(i)}}{a_\lambda^{(0)}}$$

Concatenating the results from (a), (b) and (c) then results in expressions (6.51) for the variational free energy.

□

The VB algorithm derived in this example is shown in pseudocode form in Table 3.

% Initialization of the variational parameters and variational free energy

$m_\mu^{(0)} \coloneqq m_\mu^0$

$s_\mu^{2\,(0)} \coloneqq s_\mu^{2\,0}$

$a_\lambda^{(0)} \coloneqq a_\lambda^0$

$b_\lambda^{(0)} \coloneqq b_\lambda^0$

% evaluation of the initial variational free energy

$\mathcal{F}^{(0)} \coloneqq \mathcal{F}\left(y, N, m_\mu^{(0)}, s_\mu^{2\,(0)}, m_\mu^{(0)}, s_\mu^{2\,(0)}, a_\lambda^{(0)}, b_\lambda^{(0)}, a_\lambda^{(0)}, b_\lambda^{(0)}\right)$

% VB iterations

for $i = 0,1,2,\ldots$until convergence do

% E-Step

$m_\mu^{(i+1)} \coloneqq \frac{m_\mu^{(i)} + s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}}$

$s_\mu^{2\,(i+1)} \coloneqq \frac{s_\mu^{2\,(i)}}{1 + N s_\mu^{2\,(i)} a_\lambda^{(i)} b_\lambda^{(i)}}$

% M-Step

$a_\lambda^{(i+1)} \coloneqq \frac{N}{2} + a_\lambda^{(i)}$

$b_\lambda^{(i+1)} \coloneqq \left(\frac{1}{b_\lambda^{(i)}} + \frac{1}{2}\left(\sum_{n=1}^N y_n^2 - 2\sum_{n=1}^N y_n\, m_\mu^{(i+1)} + N\left(\left(m_\mu^{(i+1)}\right)^2 + s_\mu^{2\,(i+1)}\right)\right)\right)^{-1}$

% Free Energy Evaluation

$\mathcal{F}^{(i+1)} \coloneqq \mathcal{F}\left(y, N, m_\mu^{(i+1)}, s_\mu^{2\,(i+1)}, m_\mu^{(0)}, s_\mu^{2\,(0)}, a_\lambda^{(i+1)}, b_\lambda^{(i+1)}, a_\lambda^{(0)}, b_\lambda^{(0)}\right)$

end

% VB posterior unobserved variable distribution and log model evidence approximation

$p(\mu, \lambda|y)_{VB} \coloneqq N\left(\mu; m_\mu^{(i+1)}, s_\mu^{2\,(i+1)}\right) \cdot G\left(\lambda; a_\lambda^{(i+1)}, b_\lambda^{(i+1)}\right)$

$\ln p(y)_{VB} \coloneqq \mathcal{F}^{(i+1)}$

Table 3. Pseudocode for the VB Algorithm for the univariate Gaussian.

7 The Kalman-Rauch-Tung-Striebel Smoothing Algorithm

For LGSSMs, the evaluation of conditional distributions over the latent variables $x_{1:T}$ given an observed data sequence $y^*_{1:T}$ and a known parameter $\theta^{(i)}$ corresponds to the "state space model inference" problem for which a variety of solutions exist in the literature (for a comprehensive review see e.g. (Briers, Doucet, & Maskell, 2010)). The most popular solution is perhaps the KRTS smoother, which represents the combination of the classic Kalman Filter for state space models (Kalman, 1960) with a backward algorithm (Johari, Desabrais, Rauch, Striebel, & Tung, 1965). In brief, the KRTS smoother is a recursive algorithm that yields the expectation and covariance parameters of the conditional marginal latent variable distributions in matrix product form.

Before proceeding, it is helpful to differentiate the terms "filtering" and "smoothing". As stated above, the aim of inference for LGSSMs is to derive probabilistic conclusions about the states of a hidden variable $x_{1:T}$ given a realization sequence of the observed variables set $y_{1:T} = y^*_{1:T}$. For each $x_t,\ t = 1, \dots, T$, these conclusions can in principle be derived from the joint distribution of the LGSSM $p(x_{1:T}, y_{1:T})$ as specified in equation (9) of the main tutorial by appropriately conditioning on the remaining variables. Depending on whether in the conditional marginal distribution $p(x_t|y_{1:k})$ the value of k is smaller than, equal to, or larger than the value of t, inference of the distribution $p(x_t|y_{1:k})$ takes different forms and is labeled differently. Specifically, for $k < t$, inferring $p(x_t|y_{1:k})$ is called "prediction," and will not be considered here. For $k = t$, inferring $p(x_t|y_{1:t})$ is referred to as "filtering." As will be seen below, filtering also forms the basis for the case of $k > t$. Inferring $p(x_t|y_{1:T})$ is known as (fixed interval) "smoothing." Importantly, in addition to the information available from observing $y_{1:t}$ as in filtering, smoothing also takes into account the evolution of the observations after x_t obtained a specific hidden state. Intuitively, the smoothed conditional marginal distribution $p(x_t|y_{1:T})$ is hence more veridical than the filtered conditional marginal distribution $p(x_t|y_{1:t})$.

In the following discussion, we omit the superscript (i) from the parameter $\theta^{(i)}$ for notational brevity, keeping in mind that in the context of the exact VML-EM the parameter θ correspond to the parameter estimate at a given iteration of the algorithm.

Denoting the dimensionality of the latent variable by $k \in \mathbb{N}$ and the dimensionality of the observed variable by $p \in \mathbb{N}$, expressions (5) of the main tutorial is written more generally as

$$x_t = Ax_{t-1} + \varepsilon_t \ \left(x_t \in \mathbb{R}^k, A \in \mathbb{R}^{k\times k}, \varepsilon_t \sim N(\varepsilon_t; 0, \Sigma_x), 0 \in \mathbb{R}^k, \Sigma_x \in \mathbb{R}^{k\times k}, \Sigma_x \succ 0\right) \tag{7.1}$$

$$y_t = Bx_t + \eta_t \quad (y_t \in \mathbb{R}^p, B \in \mathbb{R}^{p\times k}, \eta_t \sim N(\eta_t; 0, \Sigma_y), 0 \in \mathbb{R}^p, \Sigma_y \in \mathbb{R}^{p\times p}, \Sigma_y \succ 0) \tag{7.2}$$

for $t = 2, \dots, T$ and with the assumption of independent dynamics and observation noise, that is, $\mathbb{C}(\varepsilon_t, \eta_t) = 0 \in \mathbb{R}^{k\times p}$. It is central for the understanding of LGSSMs that these two equations specify a multivariate joint Gaussian distribution over the latent variables $x_2, \dots, x_T$ and the observed variables and $y_2, \dots, y_T$. Specifically, due to the Gaussian properties of ε_t and η_t, equations (71) and (7.2) specify the following conditional probability distributions over state and observations variables x_t and y_t, respectively

$$p(x_t|x_{t-1}) = N(x_t; Ax_{t-1}, \Sigma_x) \text{ and } p(y_t|x_t) = N(y_t; Bx_t, \Sigma_y),\ t = 2, \dots, T \tag{7.3}$$

The specification of the joint distribution for the LGSSM is completed by a further Gaussian distribution over x_1 with expectation parameter $\mu_1 \in \mathbb{R}^k$ and covariance $\Sigma_1 \in \mathbb{R}^{k\times k}, \Sigma_x \succ 0$,

$$p(x_1) = N(x_1;\ \mu_1, \Sigma_1) \tag{7.4}$$

and the identification of

$$p(y_1|x_1) \coloneqq N(y_1;\ Bx_1, \Sigma_y) \tag{7.5}$$

Using the "colon notation" abbreviations

$$x_{1:T} := (x_1, \dots, x_T) \text{ and } y_{1:T} := (y_1, \dots, y_T) \tag{7.6}$$

then allows for writing the joint distribution over latent and observed variables as

$$p_\theta(x_{1:T}, y_{1:T}) = p_\theta(x_1) \prod_{t=2}^{T} p_\theta(x_t|x_{t-1}) \prod_{t=1}^{T} p_\theta(y_t|x_t) \tag{7.7}$$

Here, the subscript θ refers to the parameter[4] of this distribution given by $\theta := \{\mu_1, \Sigma_1, A, \Sigma_x, B, \Sigma_y\}$. As discussed in the tutorial, expression (5) states that the joint distribution over all variables $x_{1:T}$ and $y_{1:T}$ of the LGSSM is given by the product of Gaussian marginal and conditional distributions over x_1 and $x_{2:T}, y_{1:T}$, respectively. Because the product of Gaussian probability densities is again a Gaussian probability densitiy, (7.7) hence specifies a Gaussian joint distribution by means of its factorization properties.

Kalman Filtering – Inferring $p_\theta(x_t|y_{1:t})$

Inferring a probability distribution over the hidden variable x_t given all observations up to y_t, i.e., $p_\theta(x_t|y_{1:t})$ is performed readily due to the analytical properties of Gaussians distributions. As the joint distribution over $x_{1:T}$ and $y_{1:T}$ is Gaussian, all marginal and conditional marginal distributions are also Gaussian distributions. Further, as Gaussian distributions are characterized by their first two central moments, i.e., their expectation and covariance parameters, it is only these parameters that have to be inferred to fully characterize the filter distribution of interest

$$p_\theta(x_t|y_{1:t}) = N(x_t; \mu_{x_t|y_{1:t}}, \Sigma_{x_t|y_{1:t}}) \tag{7.8}$$

The aim of the following section is to find a recursive expression for the conditional expectation and covariance parameters $\mu_{x_t|y_{1:t}} \in \mathbb{R}^k$ and $\Sigma_{x_t|y_{1:t}} \in \mathbb{R}^{k\times k}, \Sigma_{x_t|y_{1:t}} \succ 0$ of (7.8) in terms of the parameters of the "preceding distribution"

$$p_\theta(x_{t-1}|y_{1:t-1}) = N(x_{t-1}; \mu_{x_{t-1}|y_{1:t-1}}, \Sigma_{x_{t-1}|y_{1:t-1}}) \tag{7.9}$$

Formally, this can be achieved in two steps[5]: first, by finding the parameters of

$$p_\theta(x_t, y_t|y_{1:t-1}) = N\left(\binom{x_t}{y_t}; \mu_{x_t,y_t|y_{1:t-1}}, \Sigma_{x_t,y_t|y_{1:t-1}}\right) \tag{7.10}$$

in terms of the parameters of the distribution $p_\theta(x_{t-1}, y_{t-1}|y_{1:t-1})$, and second, by conditioning $p_\theta(x_t, y_t|y_{1:t-1})$ on y_t, yielding the parameters of the distribution of interest (7.8). The first step can be achieved by considering the partitioning of the expectation and covariance parameters of $p_\theta(x_t, y_t|y_{1:t-1})$, which are given by

$$\mu_{x_t,y_t|y_{1:t-1}} = \binom{\mu_{x_t|y_{1:t-1}}}{\mu_{y_t|y_{1:t-1}}} \in \mathbb{R}^{k+p} \tag{7.11}$$

and

$$\Sigma_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} \mathbb{C}(x_t, x_t|y_{1:t-1}) & \mathbb{C}(x_t, y_t|y_{1:t-1}) \\ \mathbb{C}(y_t, x_t|y_{1:t-1}) & \mathbb{C}(y_t, y_t|y_{1:t-1}) \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)} \tag{7.12}$$

Recursive expressions of (7.11) and (7.12) in terms of the parameters of $p_\theta(x_{t-1}, y_{t-1}|y_{1:t-1})$ can be obtained by means of the linear transformation theorem for Gaussian distributions, yielding (see below for details)

[4]Although θ obviously represents a "set of parameters", or, if the elements of this were concatenated into a column vector, a "parameter vector", we refer to it simply as "parameter". This practice eases the linguistic transition between models comprising a single parameter and those comprising multiple parameters.

[5] In the derivation of the Kalman-filter equations, we follow (David Barber, 2012)

$$\mu_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} A\mu_{x_{t-1}|y_{1:t-1}} \\ BA\mu_{x_{t-1}|y_{1:t-1}} \end{pmatrix} \in \mathbb{R}^{k+p} \tag{7.13}$$

and

$$\Sigma_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x & B^T\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right) \\ B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)^T & B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T + \Sigma_y \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)} \tag{7.14}$$

○ ***Derivation of equations (7.13) and (7.14)***

The aim of this section is to show how recursive expressions for the parameters of the joint Gaussian distribution $p_\theta(x_t, y_t|y_{1:t-1})$ (denoted by $\mu_{x_t,y_t|y_{1:t-1}}$ and $\Sigma_{x_t,y_t|y_{1:t-1}}$) can be obtained in terms of the parameters of the "preceding" joint distribution $p_\Theta(x_{t-1}, y_{t-1}|y_{1:t-1})$ (denoted by $\mu_{x_{t-1},y_{t-1}|y_{1:t-1}}$ and $\Sigma_{x_{t-1},y_{t-1}|y_{1:t-1}}$). In the main text, it was noted that the parameters $\mu_{x_t,y_t|y_{1:t-1}}$ and $\Sigma_{x_t,y_t|y_{1:t-1}}$ partition according to

$$\mu_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} \mu_{x_t|y_{1:t-1}} \\ \mu_{y_t|y_{1:t-1}} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(x_t|y_{1:t-1}) \\ \mathbb{E}(y|y_{1:t-1}) \end{pmatrix} \in \mathbb{R}^{k+p} \tag{7.15}$$

and

$$\Sigma_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} \Sigma_{x_t,x_t|y_{1:t-1}} & \Sigma_{x_t,y_t|y_{1:t-1}} \\ \Sigma_{y_t,x_t|y_{1:t-1}} & \Sigma_{y_t,y_t|y_{1:t-1}} \end{pmatrix} = \begin{pmatrix} \mathbb{C}(x_t, x_t|y_{1:t-1}) & \mathbb{C}(x_t, y_t|y_{1:t-1}) \\ \mathbb{C}(y_t, x_t|y_{1:t-1}) & \mathbb{C}(y_t, y_t|y_{1:t-1}) \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)} \tag{7.17}$$

Based on the Gaussian linear transformation theorem, it is acknowledged, that if $p_\theta(x_{t-1}, y_{t-1}|y_{1:t-1})$, then $p_\theta(x_t, y_t|y_{1:t-1})$ is a Gaussian distribution as well, and inference on the distribution $p_\theta(x_{t-1}, y_{t-1}|y_{1:t-1})$ is achieved by inference on its expectation and covariance parameter components.

The first component of $\mu_{x_t,y_t|y_{1:t-1}}$ represents the expectation of the random vector x_t conditioned on the observations $y_{1:t-1}$. Based on the dynamics equation of the LGSSM it is thus obtained by a linear transformation of the expectation of the random vector x_{t-1} conditioned on the observations $y_{1:t-1}$ under the transition matrix A and additive zero-mean noise ε_t, in other words

$$\begin{aligned} \mu_{x_t|y_{1:t-1}} &= \mathbb{E}(x_t|y_{1:t-1}) \\ &= \mathbb{E}(Ax_{t-1} + \varepsilon_t|y_{1:t-1}) \\ &= \mathbb{E}(Ax_{t-1}|y_{1:t-1}) \\ &= A\mathbb{E}(x_{t-1}|y_{1:t-1}) \\ &= A\mu_{x_{t-1}|y_{1:t-1}} \end{aligned} \tag{7.16}$$

Likewise, the expectation of the observation y_t conditioned on the observations $y_{1:t-1}$ is given by the expectation of the linear transformation of x_t under the emission matrix B and additive zero-mean noise η_t; and further, x_t is itself given by the linear transformation of x_{t-1} under A, andd additive zero-mean noise ε_t in other words

$$\begin{aligned} \mu_{y_t|y_{1:t-1}} &= \mathbb{E}(y_t|y_{1:t-1}) \\ &= \mathbb{E}(Bx_t + \eta_t|y_{1:t-1}) \\ &= \mathbb{E}(BAx_{t-1} + \varepsilon_t|y_{1:t-1}) \\ &= BA\mathbb{E}(x_{t-1}|y_{1:t-1}) \end{aligned} \tag{7.17}$$

$$= BA\mu_{x_{t-1}|y_{1:t-1}}$$

Concatenation then yields the desired recursive expression for the expectation parameter of $p_\theta(x_t, y_t|y_{1:t-1})$

$$\mu_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} A\mu_{x_{t-1}|y_{1:t-1}} \\ BA\mu_{x_{t-1}|y_{1:t-1}} \end{pmatrix} \tag{7.18}$$

The first component of $\Sigma_{x_t,y_t|y_{1:t-1}}$ represents the covariance of the random vector x_t with itself conditioned on $y_{1:t-1}$, $\mathbb{C}(x_t, x_t|y_{1:t-1})$. As x_t is given as the linear transformation of x_{t-1} under additive zero mean noise ε_t with covariance matrix Σ_x, the Gaussian linear transformation theorem (conditioned on $y_{1:t-1}$) applies. The expression for the covariance of x_t with x_t in terms of the covariance of x_{t-1} with x_{t-1} is thus given by

$$\mathbb{C}(x_t, x_t|y_{1:t-1}) = A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x \tag{7.19}$$

Evaluating the covariance of y_t with y_t in terms of the covariance of y_{t-1} with y_{t-1} yields

$$\mathbb{C}(y_t, y_t|y_{1:t-1}) = \mathbb{C}(Bx_t + \eta_t\,, Bx_t + \eta_t|y_{1:t-1}) \tag{7.20}$$

$$= \mathbb{C}(Bx_t\,, Bx_t|y_{1:t-1}) + \Sigma_y$$

$$= B\mathbb{C}(x_t\,, x_t|y_{1:t-1})B^T + \Sigma_y$$

$$= B(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x)B^T + \Sigma_y$$

Finally, evaluation of the covariance of x_t with y_t n terms of the covariance of x_{t-1} with y_{t-1} yields

$$\mathbb{C}(x_t, y_t|y_{1:t-1}) = \mathbb{C}(Ax_{t-1} + \varepsilon_t\,, B(Ax_{t-1} + \varepsilon_t) + \eta_t|y_{1:t-1}) \tag{7.21}$$

$$= \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_{\mathrm{x}}\right)B^T$$

and hence also

$$\mathbb{C}(y_t, x_t|y_{1:t-1}) = \left(\mathbb{C}(x_t, y_t|y_{1:t-1})\right)^T \tag{7.22}$$

$$= B^T\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)$$

$$= \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B$$

Concatenation then yields the desired recursive expression for the covariance parameter of $p_\theta(x_t, y_t|y_{1:t-1})$:

$$\Sigma_{x_t,y_t|y_{1:t-1}} = \begin{pmatrix} A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x & B^T\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right) \\ B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)^T & B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T + \Sigma_y \end{pmatrix} \in \mathbb{R}^{(p+k)\times(p+k)} \tag{7.23}$$

□

The second step can be achieved by capitalizing on the Gaussian conditioning theorem, resulting in

$$\mu_{x_t|y_{1:t}} = A\mu_{x_{t-1}|y_{1:t-1}} + \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T\left(B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T + \Sigma_y\right)^{-1}(y_t - BA\mu_{x_{t-1}|y_{1:t-1}})$$

(7.24)

and

$$\Sigma_{x_t|y_{1:t}} = \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right) - \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T\left(B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T + \Sigma_y\right)^{-1}B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)^T$$

(7.25)

○ ***Derivation of equations (7.24) and (7.25)***

Based on the recursive expressions of the parameters of $p_\theta\,(x_t, y_t|y_{1:t-1})$ in terms of the parameters of $p_\theta\,(x_{t-1}, y_{t-1}|y_{1:t-1})$, the recursive expressions for the parameters of $p_\theta\,(x_t|y_{1:t})$ by means of applying the Gaussian conditioning theorem. Specifically, by setting

$$z \coloneqq \binom{x_t}{y_t}, x_t \in \mathbb{R}^p, y_t \in \mathbb{R}^k \tag{7.26}$$

and using

$$p(x_t|y_t) = N\left(x; \mu_{x_t|y_t}, \Sigma_{x_t|y_t}\right) \tag{7.29}$$

where

$$\mu_{x_t|y_t} = \mu_{x_t} + \Sigma_{x_t y_t}\Sigma_{y_t y_t}^{-1}(y_t - \mu_{y_t}) \text{ and } \Sigma_{x_t|y_t} = \Sigma_{x_t x_t} - \Sigma_{x_t y_t}\Sigma_{y_t y_t}^{-1}\Sigma_{y_t x_t} \tag{7.27}$$

yields by substitution of the respective expresssions (conditioned on $y_{1:t-1}$):

$$\mu_{x_t|y_{1:t}} = A\mu_{x_{t-1}|y_{1:t-1}} + \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T\left(B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T + \Sigma_y\right)^{-1}\left(y_t - BA\mu_{x_{t-1}|y_{1:t-1}}\right)$$

(7.28)

and

$$\Sigma_{x_t|y_{1:t}} = \left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right) - \left(\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T\left(B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)B^T + \Sigma_y\right)^{-1}B\left(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x\right)\right)$$

(7.29)

□

The expressions for $\mu_{x_t|y_{1:t}}$ and $\Sigma_{x_t|y_{1:t}}$ above can be formulated more succinctly by setting

$$P \coloneqq A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x \in \mathbb{R}^{k\times k} \tag{7.30}$$

yielding

$$\mu_{x_t|y_{1:t}} = A\mu_{x_{t-1}|y_{1:t-1}} + PB^T\left(BPB^T + \Sigma_y\right)^{-1}(y_t - BA\mu_{x_{t-1}|y_{1:t-1}}) \tag{7.31}$$

and

$$\Sigma_{x_t|y_{1:t}} = P - PB^T\left(BPB^T + \Sigma_y\right)^{-1}BP \tag{7.32}$$

Finally, by defining the so-called Kalman Gain Matrix

$$K \coloneqq PB^T\left(BPB^T + \Sigma_y\right)^{-1} \in \mathbb{R}^{k\times p} \tag{7.33}$$

the parameters may equivalently be expressed as

$$\mu_{x_t|y_{1:t}} = A\mu_{x_{t-1}|y_{1:t-1}} + K\big(y_t - BA\mu_{x_{t-1}|y_{1:t-1}}\big) \tag{7.34}$$

and

$$\Sigma_{x_t|y_{1:t}} = P - KBP = (I - KB)P \tag{7.35}$$

To avoid that the inferred covariance $\Sigma_{x_t|y_{1:t}}$, which is given by the difference of two positive-definite matrices in (7.35) and, hence, is itself positive-definite, becomes nonpositive-definite due to numerical rounding errors, Bucy and Joseph (Bucy & Joseph, 1987) proposed the following modified update rule, which includes the formation of a positive-definite matrix on every inference step, which is derived below:

$$\Sigma_{x_t|y_{1:t}} = (I - KB)P = (I - KB)P(I - KB)^T + K\Sigma_y K^T \tag{7.36}$$

○ **Derivation of equation (7.36)**

We show the equivalence of the expressions

$$(I - KB)P(I - KB)^T + K\Sigma_{\mathrm{y}}\, K^T \text{ and } (I - KB)P \tag{7.37}$$

for the filtered covariance parameter $\Sigma_{x_t|y_{1:t}}$, where the latter representation has the benefit of involving the computation of a positive definit matrix at every recursion step.

$$\begin{aligned}(I - KB)P(I - KB)^T + K\Sigma_y\, K^T &= (I - KB)P(I - B^TK^T) + K\Sigma_y\, K^T \\ &= (I - KB)P - (I - KB)PB^TK^T + K\Sigma_y\, K^T \\ &= (I - KB)P - PB^TK^T + KBPB^TK^T + K\Sigma_y\, K^T \\ &= (I - KB)P - PB^TK^T + K\big(BPB^T + \Sigma_y\big)K^T \\ &= (I - KB)P - PB^TK^T + PB^T\left(BPB^T + \Sigma_y\right)^{-1}\big(BPB^T + \Sigma_y\big)K^T \\ &= (I - KB)P - PB^TK^T + PB^TK^T \\ &= (I - KB)P\end{aligned} \tag{7.38}$$

□

In summary, based on the initialization

$$\mu_{x_1|y_1} \coloneqq \mu_1^{(i)} \text{ and } \Sigma_{x_1|y_1} \coloneqq \Sigma_1^{(i)} \tag{7.39}$$

the parameters of the distributions $p_\theta(x_t|y_{1:t})$ can successively be obtained by using equations (7.24) and (7.25) on the ith iteration of the exact VML-EM E-Step. However, as discussed in the main tutorial, the exact VML-EM requires conditional marginal distributions of the form $p_\theta(x_t|y_{1:T})$. Parameters of these distributions can be obtained by augmenting the Kalman filter recursions with a backward recursions starting from $p_\theta(x_T|y_T)$ as discussed next.

KRTS Smoothing – Inferring $p_\theta(x_t|y_{1:T})$

Like the conditional marginal distribution $p_\theta(x_t|y_{1:t})$ the conditional marginal $p_\theta(x_t|y_{1:T})$ of an LGSSM is a normal distribution and, hence, can be characterized by its first two central moments $\mu_{x_t|y_{1:T}}$ and $\Sigma_{x_t|y_{1:T}}$. In analogy with (7.9) and (7.10), the aim of the following section is to derive the parameters of

$$p_\theta(x_t|y_{1:T}) = N(x_t; \mu_{x_t|y_{1:T}}, \Sigma_{x_t|y_{1:T}}) \tag{7.40}$$

in terms of the parameters of the "succeeding distribution"

$$p_\theta(x_{t+1}|y_{1:T}) = N(x_{t+1}; \mu_{x_{t+1}|y_{1:T}}, \Sigma_{x_{t+1}|y_{1:T}}) \tag{7.41}$$

A backward recursion starting from the final result of the Kalman Filter recursion $p_\theta(x_T|y_{1:T})$ may be derived by noting that for the LGSSM the joint distribution over temporally adjacent latent variables conditioned on $y_{1:T}$ may be written as (see below for details). In the derivation of the KRTS update equations, we follow (Yu, Shenoy, & Sahani, 2004).

$$p_\theta(x_t, x_{t+1}|y_{1:T}) = \frac{p_\theta(x_{t+1}|x_t)p_\theta(x_t|y_{1:t})p_\theta(x_{t+1}|y_{1:T})}{p_\theta(x_{t+1}|y_{1:t})} \tag{7.42}$$

○ **Derivation of equation (7.42)**
We first note that

$$p_\theta(x_t, x_{t+1}|y_{1:T}) = p_\theta(x_t|x_{t+1}, y_{1:T})p_\theta(x_{t+1}|y_{1:T}) \tag{7.43}$$

We next note that x_t is independent of y_{t+1} given x_{t+1}, i.e. given x_{t+1}, x_t and y_{t+1} are conditionally independent, and one may write

$$p_\theta(x_t|x_{t+1}, y_{1:T}) = p_\theta(x_t|x_{t+1}, y_{1:t}) \tag{7.44}$$

The conditional independence of x_t and $y_{t+1:T}$ given x_{t+1} may be derived by the using the d-separation criterion in the graphical model framework (p. 378 in (Bishop, 2007)): because every path from x_t to $y_{t+1:T}$ is blocked by the node x_{t+1}, i.e. the arrows on the respective paths meet head-to-tail at the node x_{t+1} (see Figure 3A of the main tutorial), x_t is conditional independent of $y_{t+1:T}$ given x_{t+1}. In other words

$$x_t \perp y_{k+1}|x_{t+1} \text{ for } k = t, t+1, \dots, T-1 \tag{7.45}$$

We hence have

$$p_\theta(x_t, x_{t+1}|y_{1:T}) = p_\theta(x_t|x_{t+1}, y_{1:t})p_\theta(x_{t+1}|y_{1:T}) \tag{7.46}$$

□

Considering the first term on the right-hand side of the above, we have, again with the factorization properties of the LGSSM

$$\begin{aligned} p_\theta(x_t|x_{t+1}, y_{1:t}) &= \frac{p_\theta(x_t, x_{t+1}|y_{1:t})}{p_\theta(x_{t+1}|y_{1:t})} \\ &= \frac{p_\theta(x_t|y_{1:t})p_\theta(x_{t+1}|x_t, y_{1:t})}{p_\theta(x_{t+1}|y_{1:t})} \\ &= \frac{p_\theta(x_t|y_{1:t})p_\theta(x_{t+1}|x_t)}{p_\theta(x_{t+1}|y_{1:t})} \end{aligned} \tag{7.47}$$

where the conditional independence of x_{t+1} and $y_{1:t}$ given x_t follows from

$$\begin{aligned} p_\theta(x_{t+1}, y_{1:t}|x_t) &= \frac{p_\theta(x_t, x_{t+1}, y_{1:t})}{p_\theta(x_t)} \\ &= \frac{p_\theta(x_t)p_\theta(x_{t+1}|x_t)p_\theta(y_{1:t-1}, y_t|x_t)}{p_\theta(x_t)} \\ &= p_\theta(x_{t+1}|x_t)p_\theta(y_{1:t}|x_t) \end{aligned} \tag{7.48}$$

Taking logarithms and substituting for the distributions on the right hand side of (7.48) then results in

$$\ln p_\theta(x_t, x_{t+1}|y_{1:T}) = -\frac{1}{2}(x_{t+1} - Ax_t)^T \Sigma_x^{-1}(x_{t+1} - Ax_t) \tag{7.49}$$

$$-\frac{1}{2}\left(x_t-\mu_{x_t|y_{1:t}}\right)^T\left(\Sigma_{x_t|y_{1:t}}\right)^{-1}\left(x_t-\mu_{x_t|y_{1:t}}\right)$$

$$-\frac{1}{2}\left(x_{t+1}-\mu_{x_{t+1}|y_{1:T}}\right)^T\left(\Sigma_{x_{t+1}|y_{1:T}}\right)^{-1}\left(x_t-\mu_{x_{t+1}|y_{1:T}}\right)$$

$$+\frac{1}{2}\left(x_{t+1}-\mu_{x_{t+1}|y_{1:t}}\right)^T\left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1}\left(x_t-\mu_{x_{t+1}|y_{1:t}}\right)+C$$

where $C \in \mathbb{R}$ denotes a a constant independent of x_t and x_{t+1}. Rewriting the above in terms of first and second order in x_t and x_{t+1} and comparing these to a standard Gaussian distribution partition allows for inferring the entries of the conditional covariance matrix $\Sigma_{x_t,x_{t+1}|y_{1:T}}$. This conditional covariance matrix partitions according to

$$\Sigma_{x_t,x_{t+1}|y_{1:T}}=\begin{pmatrix}\Sigma_{x_t|y_{1:T}} & \Sigma_{x_t,x_{t+1}|y_{1:T}}\\ \Sigma_{x_{t+1},x_t|y_{1:T}} & \Sigma_{x_{t+1}|y_{1:T}}\end{pmatrix}=\begin{pmatrix}\mathbb{C}(x_t,x_t|y_{1:T}) & \mathbb{C}(x_t,x_{t+1}|y_{1:T})\\ \mathbb{C}(x_t,x_{t+1}|y_{1:T}) & \mathbb{C}(x_{t+1},x_{t+1}|y_{1:T})\end{pmatrix}\in\mathbb{R}^{2k\times 2k} \tag{7.50}$$

The diagonal entries of $\Sigma_{x_t,x_{t+1}|y_{1:T}}$ in turn allow for inferring the following recursive KRTS update equations (see below for details)

$$\mu_{x_t|y_{1:T}}=\mu_{x_t|y_{1:t}}+\Sigma_{x_t|y_{1:t}}A^T\left(A\Sigma_{x_t|y_{1:t}}A^T+\Sigma_x\right)^{-1}\left(\mu_{x_{t+1}|y_{1:T}}-A\mu_{x_t|y_{1:t}}\right) \tag{7.51}$$

and

$$\Sigma_{x_t|y_{1:T}}=\Sigma_{x_t|y_{1:t}}+\left(\Sigma_{x_t|y_{1:t}}A^T\left(A\Sigma_{x_t|y_{1:t}}A^T+\Sigma_x\right)^{-1}\right)\left(\Sigma_{x_{t+1}|y_{1:T}}-\left(A\Sigma_{x_t|y_{1:t}}A^T+\Sigma_x\right)\right)\left(\Sigma_{x_t|y_{1:t}}A^T\left(A\Sigma_{x_t|y_{1:t}}A^T+\Sigma_x\right)^{-1}\right)^T \tag{7.52}$$

Notably, these recursive backward expressions for the smoothed expectation $\mu_{x_t|y_{1:T}}$ and covariance parameters $\Sigma_{x_t|y_{1:T}}$ contain only the filtered conditional expectation $\mu_{x_t|y_{1:t}}$, the filtered conditional covariance $\Sigma_{x_t|y_{1:t}}$, the transition matrix A, and the state noise covariance Σ_x. The ensuing algorithm is referred to as KRTS smoother.

○ **Derivations of of equations (7.51) and (7.52)**

As outlined above, we derive the KRTS update equations by means of the conditional covariance matrix $\Sigma_{x_t,x_{t+1}|y_{1:T}}$ following the approach of (Yu, Shenoy, & Sahani, 2004). This approach capitalizes on the following two matrix identities

$$(A+UCV)^{-1}=A^{-1}-A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \tag{7.53}$$

and

$$(I-(A+B)^{-1}A)B^{-1}=(A+B)^{-1} \tag{7.54}$$

for appropriately specified matrices A, B, C, U and V. (7.53) is known as the "Woodbury identity" or "Matrix Inversion Lemma" and its verification is left to the interested reader. (7.54) is verified as

$$\begin{aligned}&(A+B)^{-1}(A+B)=I\\ \Leftrightarrow\ &(A+B)^{-1}A+(A+B)^{-1}B=I\\ \Leftrightarrow\ &(A+B)^{-1}B=I-(A+B)^{-1}A\\ \Leftrightarrow\ &(A+B)^{-1}=(I-(A+B)^{-1}A)B^{-1}\\ \Leftrightarrow\ &(I-(A+B)^{-1}A)B^{-1}=(A+B)^{-1}\end{aligned} \tag{7.55}$$

Further, the approach of (Yu, Shenoy, & Sahani, 2004) uses the formulation of the quadratic form in the exponent of multivariate Gaussian distributions as introduced in Supplement Section 2 and notated here[6] as

$$\ln p(z_1, z_2) = -\frac{1}{2}\begin{pmatrix} z_1 - \mu_1 \\ z_2 - \mu_2 \end{pmatrix}' \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \begin{pmatrix} z_1 - \mu_1 \\ z_2 - \mu_2 \end{pmatrix} + C \tag{7.56}$$

$$= -\frac{1}{2} z_1' \Lambda_{11} z_1 - \frac{1}{2} z_1' \Lambda_{12} z_2 - \frac{1}{2} z_2' \Lambda_{21} z_1 - \frac{1}{2} z_2' \Lambda_{22} z_2 + z_2' (\Lambda_{21} \mu_1 + \Lambda_{22} \mu_{22}) + C$$

for $p(z_1, z_2) = N\left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Lambda^{-1}\right)$ with suitably chosen vectors z_1, z_2, μ_1, μ_2 and block matrix Λ, denoting the inverse of the covariance matrix of $p(z_1, z_2)$ and $C \in \mathbb{R}$. In addition (Yu, Shenoy, & Sahani, 2004) note that for the inversion of block matrices, the following equalities hold. Let

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \coloneqq \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} \tag{7.57}$$

and

$$\Gamma_{11} \coloneqq \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} \text{ and } \Gamma_{22} \coloneqq \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} \tag{7.58}$$

Then

$$\begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Gamma_{11}^{-1} & -\Gamma_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1} \\ -\Lambda_{22}^{-1}\Lambda_{21}\Gamma_{11}^{-1} & \Gamma_{22}^{-1} \end{pmatrix} \tag{7.59}$$

$$= \begin{pmatrix} \Lambda_{11}^{-1} + \Lambda_{11}^{-1}\Lambda_{12}\Gamma_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1} & -\Gamma_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1} \\ -\Lambda_{22}^{-1}\Lambda_{21}\Gamma_{11}^{-1} & \Lambda_{22}^{-1} + \Lambda_{22}^{-1}\Lambda_{21}\Gamma_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1} \end{pmatrix}$$

Again, the verification of (7.59) is left as an exercise to the interested reader.

Based on the set of preliminaries (7.54) – (7.59), equations (7.51) and (7.52) may now be derived from (7.49) as follows. As a first step, the right hand side of (7.49) is reordered in terms of second-order terms in x_{t+1}, first-order mixed terms in x_{t+1} and x_t, second-oder terms in x_t and first-oder terms in x_t. This reordering results in

$$\ln p(x_t, x_{t+1}|y_{1:T}) = -\frac{1}{2} x_{t+1}^T \left(\Sigma_x^{-1} - \left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1} + \left(\Sigma_{x_{t+1}|y_{1:T}}\right)^{-1}\right) x_{t+1} \tag{7.60}$$

$$-\frac{1}{2} x_{t+1}^T (-\Sigma_x^{-1} A) x_t - \frac{1}{2} x_t^T (-A\Sigma_x^{-1}) x_{t+1}$$

$$-\frac{1}{2} x_t^T \left(A^T \Sigma_x^{-1} A + \left(\Sigma_{x_t|y_{1:t}}\right)^{-1}\right) x_t$$

$$+ x_t^T \left(\left(\Sigma_{x_t|y_{1:t}}\right)^{-1} \mu_{x_t|y_{1:t}}\right) + C$$

Term-by-term comparison of (7.60) and (7.56) then allows for inferring the entries of the inverse covariance matrix of $p(x_{t+1}, x_t|y_{1:T})$ to be given as

[6] In contrast to Supplement Section 2 we here use numerical indices to avoid confusion with the nomenclature for latent and observed variables in the LGSSM context.

$$\begin{pmatrix} \Sigma_{x_{t+1}|y_{1:T}} & \Sigma_{x_{t+1,t}|y_{1:T}} \\ \Sigma_{x_{t,t+1}|y_{1:T}} & \Sigma_{x_t|y_{1:T}} \end{pmatrix} = \begin{pmatrix} \Sigma_x^{-1} - \left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1} + \left(\Sigma_{x_{t+1}|y_{1:T}}\right)^{-1} & -\Sigma_x^{-1}A \\ -A\Sigma_x^{-1} & A^T\Sigma_x^{-1}A + \left(\Sigma_{x_t|y_{1:t}}\right)^{-1} \end{pmatrix}^{-1}$$

(7.61)

To obtain the covariance matrix of $p(x_{t+1}, x_t|y_{1:T})$, the block matrix on the right-hand side of (7.61) has to be inverted. To achieve this, and simultaneously derive (7.52) (Yu, Shenoy, & Sahani, 2004), first simplify Λ_{22}^{-1} and $\Lambda_{22}^{-1}\Lambda_{21}$ in (7.59) in terms of the entries in (7.61). Specifically, using the matrix inversion lemma

$$\Lambda_{22}^{-1} = \left(A^T\Sigma_x^{-1}A + \Sigma_{x_t|y_{1:T}}^{-1}\right)^{-1} = \Sigma_{x_t|y_{1:t}} - \Sigma_{x_t|y_{1:t}}A^T\left(\Sigma_{x_t|y_{1:t}+1}\right)^{-1}A\Sigma_{x_t|y_{1:t}} \tag{7.62}$$

and defining

$$J_t \coloneqq \Sigma_{x_t|y_{1:t}}A^T\left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1} \tag{7.63}$$

we obtain

$$\Lambda_{22}^{-1} = \Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t^T \tag{7.64}$$

Likewise, by means of the matrix identity

$$(I - (A + B)^{-1}A)B^{-1} = (A + B)^{-1} \tag{7.65}$$

we obtain

$$\Lambda_{22}^{-1}\Lambda_{21} = -\left(\Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t^T\right)A^T\Sigma_x^{-1} = -J_t \tag{7.66}$$

After this simplification, the block-matrix on the right-hand side of (7.61) may be inverted to obtain the entries of the covariance matrix of $p(x_{t+1}, x_t|y_{1:T})$ according to (7.52). Specifically, we have

$$\Sigma_{x_{t+1}|y_{1:T}} = \Gamma_{11}^{-1} \tag{7.67}$$

and

$$\begin{aligned} \Sigma_{x_t|y_{1:T}} &\coloneqq \Lambda_{22}^{-1} + \Lambda_{22}^{-1}\Lambda_{21}\Gamma_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1} \qquad (7.68) \\ &= (\Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t^T) + (-J_t)\Sigma_{x_{t+1}|y_{1:T}}(-J_t^T) \\ &= \Sigma_{x_t|y_{1:t}} + J_t\left(\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}}\right)J_t^T \\ &= \Sigma_{x_t|y_{1:t}} + \left(\Sigma_{x_t|y_{1:t}}A^T\left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1}\right)\left(\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}}\right)\left(\Sigma_{x_t|y_{1:t}}A^T\left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1}\right)^T \end{aligned}$$

Noting that with the Gaussian linear transformation theorem

$$\Sigma_{x_t|y_{1:t-1}} = A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x \tag{7.69}$$

and thus

$$\Sigma_{x_{t+1}|y_{1:t}} = A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x \tag{7.70}$$

we have thus (7.52), i.e. a backward recursive expression for $\Sigma_{x_t|y_{1:T}}$ in terms of the filtered covariance matrices $\Sigma_{x_t|y_{1:t}}$ and the parameters of the LGSSM

$$\Sigma_{x_t|y_{1:T}} = \Sigma_{x_t|y_{1:t}} + \left(\Sigma_{x_t|y_{1:t}} A^T \left(A\Sigma_{x_t|y_{1:t}} A^T + \Sigma_x\right)^{-1}\right)\left(\Sigma_{x_{t+1}|y_{1:T}} - \left(A\Sigma_{x_t|y_{1:t}} A^T + \Sigma_x\right)\right)\left(\Sigma_{x_t|y_{1:t}} A^T \left(A\Sigma_{x_t|y_{1:t}} A^T + \Sigma_x\right)^{-1}\right)^T \tag{7.71}$$

Finally, to find the mean, we compare the last terms in (7.56) and (7.60), i.e.

$$z_2'(\Lambda_{21}\mu_1 + \Lambda_{22}\mu_2) \text{ and } x_t^T\left(\left(\Sigma_{x_t|y_{1:t}}\right)^{-1}\mu_{x_t|y_{1:t}}\right) \tag{7.72}$$

Apparently, we have

$$\mu_1 \coloneqq \mu_{x_{t+1}|y_{1:T}} \text{ and } \mu_2 \coloneqq \mu_{x_t|y_{1:T}} \tag{7.73}$$

We hence infer

$$\Lambda_{21}\mu_{x_{t+1}|y_{1:T}} + \Lambda_{22}\mu_{x_t|y_{1:T}} = \Sigma_{x_t|y_{1:t}}^{-1}\mu_{x_t|y_{1:t}} \tag{7.74}$$

Solving for $\mu_{x_t|y_{1:T}}$, we obtain

$$\mu_{x_t|y_{1:T}} = -\Lambda_{22}^{-1}\Lambda_{21}\mu_{x_{t+1}|y_{1:T}} + \Lambda_{22}^{-1}\Sigma_{x_t|y_{1:t}}^{-1}\mu_{x_t|y_{1:t}} \tag{7.75}$$

Next, using

$$\Lambda_{22}^{-1}\Lambda_{21} = -J_t \text{ and } \Lambda_{22}^{-1} = \Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t' \tag{7.76}$$

we obtain

$$\begin{aligned}
\mu_{x_t|y_{1:T}} &= J_t\mu_{x_{t+1}|y_{1:T}} + (I - J_tA)\mu_{x_t|y_{1:t}} \\
&= J_t\mu_{x_{t+1}|y_{1:T}} + \mu_{x_t|y_{1:t}} - J_tA\mu_{x_t|y_{1:t}} \\
&= \mu_{x_t|y_{1:t}} + J_t(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}) \\
&= \mu_{x_t|y_{1:t}} + (\Sigma_{x_t|y_{1:t}}A^T\left(\Sigma_{x_{t+1}|y_{1:t}}\right)^{-1})(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}) \\
&= \mu_{x_t|y_{1:t}} + (\Sigma_{x_t|y_{1:t}}A^T\left(A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x\right)^{-1})(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}) \\
&= \mu_{x_t|y_{1:t}} + \left(\Sigma_{x_t|y_{1:t}}A^T\left(A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x\right)^{-1}\right)\left(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}\right)
\end{aligned} \tag{7.77}$$

Finally, equations (7.51) and (7.52) may be written more succinctly by defining

$$J_t \coloneqq \Sigma_{x_t|y_{1:t}}A^T\left(A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x\right)^{-1} \tag{7.78}$$

resulting in

$$\mu_{x_t|y_{1:T}} = \mu_{x_t|y_{1:t}} + J_t\left(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}\right) \tag{7.79}$$

and

$$\Sigma_{x_t|y_{1:T}} = \Sigma_{x_t|y_{1:t}} + J_t\left(\Sigma_{x_{t+1}|y_{1:T}} - \left(A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x\right)\right)J_t^T \tag{7.80}$$

This recursion is initialized as

$$\Sigma_{x_{T-1:T}|y_{1:T}} = \Sigma_{x_T|y_{1:T}} J_{T-1}^T = (I - (A\Sigma_{x_T|y_{1:T-1}} A^T + \Sigma_x) B^T \left(\Sigma_y + B(A\Sigma_{x_{T-1}|y_{1:T-1}} A^T + \Sigma_x) B^T\right)^{-1} B) A\Sigma_{x_T|y_{1:T}}$$

(7.81)

by substitution of the recursive expressions for $\Sigma_{x_T|y_{1:T}}$ and the definition of J_{T-1}. Notably, like the recursions for $\mu_{x_t|y_{1:T}}$ and $\Sigma_{x_t|y_{1:T}}$, this backward recursion contains only the filtered conditional covariance $\Sigma_{x_t|y_{1:t}}$, the transition matrix A, and the state noise covariance Σ_x. The recursion can be initialized using

$$\Sigma_{x_{T-1:T}|y_{1:T}} = \left(I - \left(A\Sigma_{x_{T-1}|y_{1:T-1}} A^T + \Sigma_x\right) B^T \left(\Sigma_y + B\left(A\Sigma_{x_{T-1}|y_{1:T-1}} A^T + \Sigma_x\right) B^T\right)^{-1} B\right) A\Sigma_{x_T|y_{1:T}}$$

(7.82)

○

To complete the current section, we demonstrate how these are obtained. We first note that, in parameterized form, the smoothed latent variable first- and second-order moment are given as

$$\langle x_t \rangle_{p_{\theta^{(i)}}(x_t|y_{1:T})} = \mu_{x_t|y_{1:T}} \text{ and } \langle x_t x_t^T \rangle_{p_{\theta^{(i)}}(x_t|y_{1:T})} = \mu_{x_t|y_{1:T}} \mu_{x_t|y_{1:T}}^T + \Sigma_{x_t|y_{1:T}} \tag{7.83}$$

and the second moment of the pairwise conditional marginal is given as

$$\langle x_{t-1} x_t^T \rangle_{p_{\theta^{(i)}}(x_{t-1}, x_t|y_{1:T})} = \mu_{x_{t-1}|y_{1:T}} \mu_{x_t|y_{1:T}}^T + \Sigma_{x_{t-1:t}|y_{1:T}} \tag{7.84}$$

The first term on the right-hand side of (7.82) capitalizes on the availability of the smoothed latent variable expectations. A recursion for the second term of (7.82) can be obtained by considering the off-diagonal elements in (5.2.24) yielding (see below for details)

$$\Sigma_{x_{t-1:t}|y_{1:T}} = \Sigma_{x_t|y_{1:t}} J_t^T \tag{7.85}$$

○ **Derivation of equation (7.80)**

According to

$$\Sigma_{x_{t,t+1}|y_{1:T}} = -\Lambda_{22}^{-1} \Lambda_{21} \Gamma_{11}^{-1} \tag{7.86}$$

and with

$$\Sigma_{x_{t+1}|y_{1:T}} = \Gamma_{11}^{-1} \text{ and } \Lambda_{22}^{-1} \Lambda_{21} = -J_t \tag{7.87}$$

we have

$$\Sigma_{x_{t:t+t}|y_{1:T}} = \Sigma_{x_{t+1}|y_{1:T}} J_t^T \Leftrightarrow \Sigma_{x_{t-1:t}|y_{1:T}} = \Sigma_{x_t|y_{1:T}} J_{t-1}^T \tag{7.88}$$

Substitution of (D.5.39) and (D.5.34) then yields

$$\Sigma_{x_{t-1:t}|y_{1:T}} = (\Sigma_{x_t|y_{1:t}} + J_t\left(\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}}\right) J_t^T) J_{t-1}^T \tag{7.89}$$

$$= \left(\Sigma_{x_t|y_{1:t}} + J_t\left(\Sigma_{x_{t+1}|y_{1:T}} - \mathrm{A}\Sigma_{x_t|y_{1:t}}\right)\right) J_{t-1}^T$$

$$= \Sigma_{x_t|y_{1:t}} \left(\left(\Sigma_{x_{t-1}|y_{1:t-1}} A^T \left(A \Sigma_{x_{t-1}|y_{1:t-1}} A^T + \Sigma_x \right)^{-1} \right) \right)$$

$$+ \left(\Sigma_{x_t|y_{1:t}} A^T \left(A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x \right)^{-1} \right) \left(\Sigma_{x_{t+1}|y_{1:T}} - \mathrm{A} \Sigma_{x_t|y_{1:t}} \right) \left(\Sigma_{x_{t-1}|y_{1:t-1}} A^T \left(A \Sigma_{x_{t-1}|y_{1:t-1}} A^T + \Sigma_x \right)^{-1} \right)^T$$

□

8 Unified Inference for LGSSMs

In contrast to the exact VML-EM algorithm, on a given iteration of the VB-EM algorithm it is not the parameters θ which are fixed to specific values, but the variational distribution $q(\theta)$ over parameters. Further, the general update rule for the VB-EM E-Step corresponds to (cf. equation (17) of the main tutorial)

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(x_{1:T}, y_{1:T}, \theta)\rangle_{q^{(i)}(\theta)}\right) \tag{8.1}$$

The inference procedures considered in Supplement Section 7, which result in the specification of $p(x_t|y_{1:T})$ and $p(x_{t-1,t}|y_{1:T})$, however, are applicable in the case of known parameter values. Intuitively, one might think that, for the E-Step of the VB-EM algorithm, the parameter expectations under the variational distributions, i.e.,$\langle\theta\rangle_{q^{(i)}(\theta)}$, may be forwarded to the KRTS smoothing algorithm. However, this is not appropriate, as one can show that for the LGSSM the expectation of the energy function under the current variational distribution over parameters $q^{(i)}(\theta)$ does not equal the energy function with averaged parameters. In other words, in general,

$$\langle \ln\ p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)} \neq \ln p\left(x_{1:T}, y_{1:T}|\langle\theta\rangle_{q^{(i)}(\theta)}\right) \tag{8.2}$$

Barber and Chiappa (D. Barber & Chiappa, 2007) have shown how, nevertheless, standard inference algorithms may be applied for the E-Step of the VB-EM algorithm. In the following, the inequality above is discussed in further and the solution provided by (D. Barber & Chiappa, 2007) is introduced.

The inequality (8.2) is summarized in (D. Barber & Chiappa, 2007) as the "mean and fluctuation decomposition theorem." Specifically, it can be shown that the expectation of the conditional energy function under $q^{(i)}(\theta)$ can be written as the sum of two terms: (1) the energy function conditioned on the parameter expectations $\langle\theta\rangle_{q^{(i)}(\theta)}$ and (2) a fluctuation term. Specifically, using the LGSSM inherent factorization of $p(x_{1:T}, y_{1:T}|\theta)$, one obtains

$$\langle \ln\ p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)} \tag{8.3}$$

$$= \langle \ln\ (p(x_1|\theta)\prod_{t=2}^{T} p(x_t|x_{t-1},\theta)\prod_{t=1}^{T} p(y_t|x_t,\theta))\rangle_{q^{(i)}(\theta)}$$

$$= \langle \ln\big(p(x_1|\mu_1,\Sigma_1)\big)\rangle_{q^{(i)}(\mu_1,\Sigma_1)} + \sum_{t=2}^{T}\langle \ln\big(p(x_t|x_{t-1},A,\Sigma_x)\big)\rangle_{q^{(i)}(A,\Sigma_x)} + \sum_{t=1}^{T}\langle \ln\left(p\big(y_t|x_t,B,\Sigma_y\big)\right)\rangle_{q^{(i)}(B,\Sigma_y)}$$

Writing out the Gaussian forms of the probability distributions involved and summarizing terms independent of the latent variables $x_{2:T}$ in a constant $C \in \mathbb{R}$ yields

$$\begin{aligned}\langle \ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)} = {} & \langle (x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\rangle_{q^{(i)}(\mu_1,\Sigma_1)} \\ & -\tfrac{1}{2}\sum_{t=2}^{T}\langle (x_t - Ax_{t-1})^T\Sigma_x^{-1}(x_t-Ax_{t-1})\rangle_{q^{(i)}(A,\Sigma_x)} \\ & -\tfrac{1}{2}\sum_{t=2}^{T}\langle (y_t - Bx_t)^T\Sigma_y^{-1}(y_t-Bx_t)\rangle_{q^{(i)}(B,\Sigma_y)} + C\end{aligned} \tag{8.4}$$

○ **Derivation of equation (8.4)**

We have

$$\langle \ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q(\theta)} \tag{8.5}$$

$$= \langle \ln p\big(x_{1:T}, y_{1:T}\big|\mu_1, A, B, \Sigma_1, \Sigma_x, \Sigma_y\big)\rangle_{q(\mu_1,A,B,\Sigma_1,\Sigma_x,\Sigma_y)}$$

$$= \langle \ln p\big(x_{1:T}, y_{1:T}\big|\mu_1, A, B, \Sigma_1, \Sigma_x, \Sigma_y\big)\rangle_{q(\mu_1,\Sigma_1)q(A,\Sigma_x)q(B,\Sigma_y)}$$

$$= \langle \ln\big(p(x_1|\mu_1,\Sigma_1)\prod_{t=2}^{T} p(x_t|A,x_{t-1},\Sigma_x)\prod_{t=1}^{T} p\big(y_t\big|B,x_t,\Sigma_y\big)\big)\rangle_{q(\mu_1,\Sigma_1)q(A,\Sigma_x)q(B,\Sigma_y)}$$

$$= \langle \ln\big(N(x_1;\mu_1,\Sigma_1)\prod_{t=2}^{T} N(x_t;Ax_{t-1},\Sigma_x)\prod_{t=1}^{T} N(y_t;Bx_t,\Sigma_y)\big)\rangle_{q(\mu_1,\Sigma_1)q(A,\Sigma_x)q(B,\Sigma_y)}$$

$$= \langle \ln\left((2\pi)^{-\frac{k}{2}}|\Sigma_1|^{-\frac{1}{2}}\exp(-\tfrac{1}{2}(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1))\right)\rangle_{q(\mu_1,\Sigma_1)q(A,\Sigma_x)q(B,\Sigma_y)}$$
$$+\langle \ln\left(\prod_{t=2}^{T}(2\pi)^{-\frac{k}{2}}|\Sigma_x|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(x_t-Ax_{t-1})^T\Sigma_x^{-1}(x_t-Ax_{t-1})\right)\right)\rangle_{q(\mu_1,\Sigma_1)q(A,\Sigma_x)q(B,\Sigma_y)}$$
$$+\langle \ln\left(\prod_{t=1}^{T}(2\pi)^{-\frac{p}{2}}\big|\Sigma_y\big|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(y_t-Bx_t)^T\Sigma_y^{-1}(y_t-Bx_t)\right)\right)\rangle_{q(\mu_1,\Sigma_1)q(A,\Sigma_x)q(B,\Sigma_y)}$$

$$= \langle \ln\left((2\pi)^{-\frac{k}{2}}|\Sigma_1|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\right)\right)\rangle_{q(\mu_1,\Sigma_1)}$$
$$+\langle \ln\left(\prod_{t=2}^{T}(2\pi)^{-\frac{k}{2}}|\Sigma_x|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(x_t-Ax_{t-1})^T\Sigma_x^{-1}(x_t-Ax_{t-1})\right)\right)\rangle_{q(A,\Sigma_x)}$$
$$+\langle \ln\left(\prod_{t=1}^{T}(2\pi)^{-\frac{p}{2}}\big|\Sigma_y\big|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(y_t-Bx_t)^T\Sigma_y^{-1}(y_t-Bx_t)\right)\right)\rangle_{q(B,\Sigma_y)}$$

$$= \langle -\tfrac{k}{2}\ln 2\pi - \tfrac{1}{2}\ln|\Sigma_1| - \tfrac{1}{2}(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\rangle_{q(\mu_1,\Sigma_1)}$$
$$+\langle -\tfrac{k(T-1)}{2}\ln 2\pi - \tfrac{T-1}{2}\ln|\Sigma_x| - \tfrac{1}{2}\sum_{t=2}^{T}(x_t-Ax_{t-1})^T\Sigma_x^{-1}(x_t-Ax_{t-1})\rangle_{q(A,\Sigma_x)}$$
$$+\langle -\tfrac{pT}{2}\ln 2\pi - \tfrac{T}{2}\ln\big|\Sigma_y\big| - \tfrac{1}{2}\sum_{t=1}^{T}(y_t-Bx_t)^T\Sigma_x^{-1}(y_t-Bx_t)\rangle_{q(B,\Sigma_y)}$$

$$= -\tfrac{k}{2}\ln 2\pi - \langle\tfrac{1}{2}\ln|\Sigma_1|\rangle_{q(\mu_1,\Sigma_1)} - \langle(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\rangle_{q(\mu_1,\Sigma_1)}$$
$$-\tfrac{k(T-1)}{2}\ln 2\pi - \langle\tfrac{T-1}{2}\ln|\Sigma_x|\rangle_{q(A,\Sigma_x)} - \tfrac{1}{2}\sum_{t=2}^{T}\langle(x_t-Ax_{t-1})^T\Sigma_x^{-1}(x_t-Ax_{t-1})\rangle_{q(A,\Sigma_x)}$$
$$-\tfrac{pT}{2}\ln 2\pi - \langle\tfrac{T}{2}\ln\big|\Sigma_y\big|\rangle_{q(B,\Sigma_y)} - \tfrac{1}{2}\sum_{t=1}^{T}\langle(y_t-Bx_t)^T\Sigma_y^{-1}(y_t-Bx_t)\rangle_{q(B,\Sigma_y)}$$

Summarizing all terms independent of $x_1, x_2, \dots, x_T$ as a constant $C \in \mathbb{R}$, one obtains

$$\begin{aligned}\langle \ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q(\theta)} = &-\langle(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\rangle_{q(\mu_1,\Sigma_1)} \\ &-\tfrac{1}{2}\sum_{t=2}^{T}\langle(x_t-Ax_{t-1})^T\Sigma_x^{-1}(x_t-Ax_{t-1})\rangle_{q(A,\Sigma_x)} \\ &-\tfrac{1}{2}\sum_{t=1}^{T}\langle(y_t-Bx_t)^T\Sigma_y^{-1}(y_t-Bx_t)\rangle_{q(B,\Sigma_y)} + C\end{aligned} \tag{8.6}$$

□

The right-hand side of (8.4) may be decomposed into the contribution of the exponent of an LGSSM with averaged parameters $\langle A\rangle_{q^{(i)}(A|\Sigma_x)}$, $\langle B\rangle_{q^{(i)}(B|\Sigma_y)}$, $\langle\Sigma_x\rangle_{q^{(i)}(\Sigma_x)}$ and $\langle\Sigma_y\rangle_{q^{(i)}(\Sigma_y)}$ and "fluctuation" terms (see below for details)

$$\langle \ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)} = \ln p\left(x_{1:T}, y_{1:T}\middle|\langle\theta\rangle_{q^{(i)}(\theta)}\right) + F_{A,\Sigma_x} + F_{B,\Sigma_y} \tag{8.7}$$

where

$$\ln p\left(x_{1:T}, y_{1:T}\middle|\langle\theta\rangle_{q^{(i)}(\theta)}\right) \coloneqq -\frac{1}{2}\sum_{t=2}^{T}\left(x_t - \langle A\rangle_{q^{(i)}(A|\Sigma_x)}x_{t-1}\right)^T \langle\Sigma_x\rangle^{-1}_{q^{(i)}(\Sigma_x)}\left(x_t - \langle A\rangle_{q^{(i)}(A|\Sigma_x)}x_{t-1}\right) \quad (8.8)$$
$$-\frac{1}{2}\sum_{t=2}^{T}\left(y_t - \langle B\rangle_{q^{(i)}(B|\Sigma_y)}x_t\right)^T \langle\Sigma_y\rangle^{-1}_{q^{(i)}(\Sigma_y)}\left(y_t - \langle B\rangle_{q^{(i)}(B|\Sigma_y)}x_t\right)$$

and

$$F_{A,\Sigma_x} \coloneqq \sum_{t=1}^{T} x_t^T\left(\langle A^T\Sigma_x^{-1}A\rangle_{q^{(i)}(A,\Sigma_x)} - \langle A^T\rangle_{q^{(i)}(A|\Sigma_x)}\langle\Sigma_x^{-1}\rangle_{q^{(i)}(\Sigma_x)}\langle A\rangle_{q^{(i)}(A|\Sigma_x)}\right)x_t \quad (8.9)$$

and

$$F_{B,\Sigma_y} \coloneqq \sum_{t=1}^{T} x_t^T\left(\langle B^T\Sigma_y^{-1}B\rangle_{q^{(i)}(B,\Sigma_y)} - \langle B^T\rangle_{q^{(i)}(B|\Sigma_y)}\langle\Sigma_y^{-1}\rangle_{q^{(i)}(\Sigma_y)}\langle B\rangle_{q^{(i)}(B|\Sigma_y)}\right)x_t \quad (8.10)$$

Given that the additional fluctuation terms F_{A,Σ_x} and F_{B,Σ_y} do not vanish in but the trivial cases, standard LGSSM inference algorithms supplied with the expected LGSSM parameters $\langle\theta\rangle_{q^{(i)}(\theta)}$ are thus not appropriate for inference in the VB-EM E-Step. In order to nevertheless use standard inference algorithms also in the Bayesian framework (and thus capitalize on the vast literature that has dealt with inference problems for LGSSMs in the past), the idea is to augment the observed variables and parameters of the "averaged LGSSM" $\langle\ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)}$ appropriately. Specifically, by applying standard inference algorithms to an augmented LGSSM log joint distribution $\ln p_{\tilde{\theta}}(x_{2:T}, \tilde{y}_{2:T})$, where $\tilde{y}_{2:T}$ denotes an augmented set of the observed variables y_t and $\tilde{\theta}$ and augmented parameters set in terms of the visible variables and parameters of $\langle\ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)}$, the latent variable distributions of the average LGSSM may be inferred. In other words, it can be shown that, for the appropriate specification of $x_{1:T}, \tilde{y}_{1:T}$ and $\tilde{\theta}$ in terms of $y_{1:T}$ and θ,

$$\ln p_{\tilde{\theta}}(x_{1:T}, \tilde{y}_{1:T}) = \langle\ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q(\theta)} \quad (8.11)$$

The required variable and parameter augmentations are specified in in the following theorem and the above equation is verified below.

Unified Inference Theorem

To infer the distribution $\ln p(x_{1:T}|y_{1:T}, \theta)$ using standard inference algorithms for the expectation of the LGSSM $\ln p(x_{1:T}, y_{1:T}|\theta)$ under the variational parameter distribution denoted by

$$\langle\ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q^{(i)}(\theta)}$$

set

$$\tilde{y}_t \coloneqq \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} \text{ and } \tilde{\theta} \coloneqq \{\tilde{\mu}_1, \tilde{\Sigma}_1\tilde{A}, \tilde{\Sigma}_x, \tilde{B}, \tilde{\Sigma}_y\}$$

where

$$\tilde{\mu}_1 \coloneqq \langle\mu_1\rangle_{q(\mu_1)} \quad \tilde{\Sigma}_1 \coloneqq \langle\Sigma_1\rangle_{q(\Sigma_1)} \quad \tilde{A} \coloneqq \langle A\rangle_{q(A|\Sigma_x)} \quad \tilde{\Sigma}_x \coloneqq \left(\langle\Sigma_x\rangle^{-1}_{q(\Sigma_x)}\right)^{-1}$$

$$\tilde{B} \coloneqq \begin{pmatrix} \langle B\rangle_{q(B)} \\ U_A \\ U_B \end{pmatrix} \quad \tilde{\Sigma}_y \coloneqq \begin{pmatrix} \left(\langle\Sigma_y\rangle^{-1}_{q(\Sigma_y)}\right)^{-1} & 0_{p\times k} & 0_{p\times k} \\ 0_{k\times p} & I_k & 0_{k\times k} \\ 0_{k\times p} & 0_{k\times k} & I_k \end{pmatrix}$$

with Cholesky decompositions $U_A$and U_B given by

$$U_A^T U_A \coloneqq \langle A^T \Sigma_x^{-1} A \rangle_{q(A,\Sigma_x)} - \langle A^T \rangle_{q(A|\Sigma_x)} \langle \Sigma_x^{-1} \rangle_{q(\Sigma_x)} \langle A \rangle_{q(A|\Sigma_x)}$$

and

$$U_B^T U_B \coloneqq \langle B^T \Sigma_y^{-1} B \rangle_{q(B,\Sigma_y)} - \langle B^T \rangle_{q(B|\Sigma_y)} \langle \Sigma_y^{-1} \rangle_{q(\Sigma_y)} \langle B \rangle_{q(B|\Sigma_y)}$$

and infer

$$ln\, p\big(x_{1:T}\big|\tilde{y}_{1:T}, \tilde{\theta}\big) = \langle ln\, p(x_{1:T}|y_{1:T}, \theta)\rangle_{q(\theta)}$$

□

- **Verification of equation (8.11)**

We verify that

$$\ln p_{\tilde{\theta}}(x_{1:T}, \tilde{y}_{1:T}) = \langle \ln p(x_{1:T}, y_{1:T}|\theta)\rangle_{q(\theta)} \tag{8.12}$$

for the augmented observed variable vector

$$\tilde{y}_t \coloneqq \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} \tag{8.13}$$

and the following augmented parameters

$$\tilde{\mu}_1 \coloneqq \mu_1,\ \tilde{\Sigma}_1 \coloneqq \Sigma_1,\ \tilde{A} \coloneqq \langle A \rangle_{q(A|\Sigma_x)}, \tilde{\Sigma}_x \coloneqq \big(\langle \Sigma_x \rangle_{q(\Sigma_x)}^{-1}\big)^{-1} \tag{8.14}$$

and

$$\tilde{B} \coloneqq \begin{pmatrix} \langle B \rangle_{q(B|\Sigma_y)} \\ U_A \\ U_B \end{pmatrix},\ \tilde{\Sigma}_y \coloneqq \begin{pmatrix} \Big(\langle \Sigma_y \rangle_{q(\Sigma_y)}^{-1}\Big)^{-1} & 0_{p\times k} & 0_{p\times k} \\ 0_{k\times p} & I_k & 0_{k\times k} \\ 0_{k\times p} & 0_{k\times k} & I_k \end{pmatrix} \tag{8.15}$$

where

$$U_A^T U_A \coloneqq \langle A^T \Sigma_x^{-1} A \rangle_{q(A,\Sigma_x)} - \langle A^T \rangle_{q(A|\Sigma_x)} \langle \Sigma_x^{-1} \rangle_{q(\Sigma_x)} \langle A \rangle_{q(A|\Sigma_x)} \tag{8.16}$$

and

$$U_B^T U_B \coloneqq \langle B^T \Sigma_y^{-1} B \rangle_{q(B,\Sigma_y)} - \langle B^T \rangle_{q(B|\Sigma_y)} \langle \Sigma_y^{-1} \rangle_{q(\Sigma_y)}\, \langle B \rangle_{q(B|\Sigma_y)} \tag{8.17}$$

by substitution. On the one hand, this verification implies that indeed the inequality (8.2) holds, while on the other other hand it simultaneously justifies the application of KRTS smoothing to the augmented LGSSM. We have

$$\ln p_{\tilde{\theta}}(x_{1:T}, \tilde{y}_{1:T}) = \ln\big(N(x_1; \tilde{\mu}_1, \tilde{\Sigma}_1) \prod_{t=2}^{T} N(x_t; \tilde{A}x_{t-1}, \tilde{\Sigma}_x) \prod_{t=1}^{T} N(\tilde{y}_t; \tilde{B}x_t, \tilde{\Sigma}_y)\big) \tag{8.18}$$

$$= -\tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}\ln\big|\tilde{\Sigma}_1\big| - \tfrac{1}{2}(x_1 - \tilde{\mu}_1)^T \tilde{\Sigma}_1^{-1}(x_1 - \tilde{\mu}_1)$$

$$-\tfrac{k(T-1)}{2}\ln 2\pi - \tfrac{T-1}{2}\ln\big|\tilde{\Sigma}_x\big| - \tfrac{1}{2}\textstyle\sum_{t=2}^{T}\Big(\big(x_t - \tilde{A}x_{t-1}\big)^T \tilde{\Sigma}_x^{-1}\big(x_t - \tilde{A}x_{t-1}\big)\Big)$$

$$-\tfrac{(p+2k)(T)}{2}\ln 2\pi - \tfrac{T}{2}\ln\big|\tilde{\Sigma}_y\big| - \tfrac{1}{2}\textstyle\sum_{t=1}^{T}\Big(\big(y_t - \tilde{B}x_t\big)^T \tilde{\Sigma}_y^{-1}\big(y_t - \tilde{B}x_t\big)\Big)$$

Summarizing all terms independent of $x_{1:T}$ into a constant, we obtain

$$\ln p_{\tilde{\theta}}(x_{1:T}, \tilde{y}_{1:T}) = -\frac{1}{2}(x_1 - \tilde{\mu}_1)^T \tilde{\Sigma}_1^{-1}(x_1 - \tilde{\mu}_1) \tag{8.19}$$

$$-\frac{1}{2}\sum_{t=2}^{T}\left(x_t - \tilde{A}x_{t-1}\right)^T \tilde{\Sigma}_x^{-1}\left(x_t - \tilde{A}x_{t-1}\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - \tilde{B}x_t\right)^T \tilde{\Sigma}_y^{-1}\left(y_t - \tilde{B}x_t\right) + C$$

Substituting of the augmented observed variable and parameters then yields

$$\ln p_{\tilde{\theta}}(x_{1:T}, \tilde{y}_{1:T}) = -\frac{1}{2}(x_1 - \mu_1)^T \Sigma_1^{-1}(x_1 - \mu_1) \tag{8.20}$$

$$-\frac{1}{2}\sum_{t=2}^{T}\left(\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)^T \langle\Sigma_x\rangle_{q(\Sigma_x)}^{-1}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T}\left(\begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} - \begin{pmatrix} \langle B\rangle_{q(B|\Sigma_y)} \\ U_A \\ U_B \end{pmatrix} x_t\right)^T \begin{pmatrix} \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} & 0_{p\times k} & 0_{p\times k} \\ 0_{k\times p} & I_k & 0_{k\times k} \\ 0_{k\times p} & 0_{k\times k} & I_k \end{pmatrix} \left(\begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} - \begin{pmatrix} \langle B\rangle_{q(B|\Sigma_y)} \\ U_A \\ U_B \end{pmatrix} x_t\right)$$

$$= -\frac{1}{2}(x_1 - \mu_1)^T \Sigma_1^{-1}(x_1 - \mu_1)$$

$$-\frac{1}{2}\sum_{t=2}^{T}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)^T \langle\Sigma_x\rangle_{q(\Sigma_x)}^{-1}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T}\begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix}^T \begin{pmatrix} \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} & 0_{p\times k} & 0_{p\times k} \\ 0_{k\times p} & I_k & 0_{k\times k} \\ 0_{k\times p} & 0_{k\times k} & I_k \end{pmatrix} \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix}$$

$$-2\begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix}^T \begin{pmatrix} \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} & 0_{p\times k} & 0_{p\times k} \\ 0_{k\times p} & I_k & 0_{k\times k} \\ 0_{k\times p} & 0_{k\times k} & I_k \end{pmatrix} \begin{pmatrix} \langle B\rangle_{q(B|\Sigma_y)} \\ U_A \\ U_B \end{pmatrix} x_t$$

$$+x_t^T \begin{pmatrix} \langle B\rangle_{q(B|\Sigma_y)} \\ U_A \\ U_B \end{pmatrix}^T \begin{pmatrix} \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} & 0_{p\times k} & 0_{p\times k} \\ 0_{k\times p} & I_k & 0_{k\times k} \\ 0_{k\times p} & 0_{k\times k} & I_k \end{pmatrix} \begin{pmatrix} \langle B\rangle_{q(B|\Sigma_y)} \\ U_A \\ U_B \end{pmatrix} x_t$$

$$= -\frac{1}{2}(x_1 - \mu_1)^T \Sigma_1^{-1}(x_1 - \mu_1)$$

$$-\frac{1}{2}\sum_{t=2}^{T}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)^T \langle\Sigma_x\rangle_{q(\Sigma_x)}^{-1}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T} y_t \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} y_t - 2y_t \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} \langle B\rangle_{q(B|\Sigma_y)}x_t$$

$$+x_t^T \langle B\rangle_{q(B|\Sigma_y)}^T \left(\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\right)^{-1} \langle B\rangle_{q(B|\Sigma_y)}x_t + x_t^T U_A^T U_A x_t + x_t^T U_B^T U_B x_t$$

$$= -\frac{1}{2}(x_1 - \mu_1)^T \Sigma_1^{-1}(x_1 - \mu_1)$$

$$-\frac{1}{2}\sum_{t=2}^{T}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)^T \langle\Sigma_x\rangle_{q(\Sigma_x)}^{-1}\left(x_t - \langle A\rangle_{q(A|\Sigma_x)}x_{t-1}\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T}\left(y_t-\langle B\rangle_{q(B|\Sigma_y)}x_t\right)^T\langle\Sigma_y\rangle_{q(\Sigma_y)}^{-1}\left(y_t-\langle B\rangle_{q(B|\Sigma_y)}x_t\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T}x_t^T\left(\langle A^T\Sigma_x^{-1}A\rangle_{q(A,\Sigma_x)}-\langle A^T\rangle_{q(A|\Sigma_x)}\langle\Sigma_x^{-1}\rangle_{q(\Sigma_x)}\langle A\rangle_{q(A|\Sigma_x)}\right)x_t$$

$$-\frac{1}{2}\sum_{t=1}^{T}x_t^T(\langle B^T\Sigma_y^{-1}B\rangle_{q(B,\Sigma_y)}-\langle B^T\rangle_{q(B|\Sigma_y)}\langle\Sigma_y^{-1}\rangle_{q(\Sigma_y)}\langle B\rangle_{q(B|\Sigma_y)})x_t$$

which yields

$$\begin{aligned}\ln p_{\tilde{\theta}}(x_{1:T},\tilde{y}_{1:T}) &= \ln p\left(x_{1:T},y_{1:T}\middle|\langle\theta\rangle_{q(\theta)}\right)\\ &+\sum_{t=1}^{T}x_t^T\left(\langle A^T\Sigma_x^{-1}A\rangle_{q(A,\Sigma_x)}-\langle A^T\rangle_{q(A|\Sigma_x)}\langle\Sigma_x^{-1}\rangle_{q(\Sigma_x)}\langle A\rangle_{q(A|\Sigma_x)}\right)x_t\\ &+\sum_{t=1}^{T}x_t^T(\langle B^T\Sigma_y^{-1}B\rangle_{q(B,\Sigma_y)}-\langle B^T\rangle_{q(B|\Sigma_y)}\langle\Sigma_y^{-1}\rangle_{q(\Sigma_y)}\langle B\rangle_{q(B|\Sigma_y)})x_t\\ &=\langle\ln p(x_{1:T},y_{1:T}|\theta)\rangle_{q(\theta)}\end{aligned} \tag{8.21}$$

□

Hence, in the augmented form, one has a standard LGSMM for which the Kalman-Rauch-Tung-Striebel algorithm can be used for inference. On the other hand, the augmented form is equivalent to the expectation of the LGSSM under the parameter distribution $q(\theta)$ as required by the VB-EM algorithm E-Step.

9 Mathematical details of the tutorial example

9.1 VB for the univariate LGSSM

We assume the following generative model

$$p\left(x_{0:T}, y_{1:T}, \mu_0, \sigma_0^2, a, \sigma_x^2, b, \sigma_y^2\right) \coloneqq \tag{9.1}$$

$$p(\mu_0)p(\sigma_0^2)p(a)p(\sigma_x^2)p(b)p\left(\sigma_y^2\right)p(x_0|\mu_0, \sigma_0^2)\prod_{t=1}^{T} p(x_t|x_{t-1}, a, \sigma_x^2)p(y_t|x_t, b, \sigma_y^2)$$

For convenience, we work with precision instead of variance parameters, i.e., we set

$$\sigma_0^2 \coloneqq \lambda_0^{-1},\ \sigma_x^2 \coloneqq \lambda_x^{-1},\ \text{and}\ \sigma_y^2 \coloneqq \lambda_y^{-1} \tag{9.2}$$

We further assume that the initial state x_0 is known as $p(x_0 = 1) = 1$, which, given

$$p(x_0|\mu_0, \sigma_0^2) = N(x_0|\mu_0, \sigma_0^2) \tag{9.3}$$

may be motivated by setting n $p(\mu_0 = 1) = 1, p(\sigma_1^2 \to 0) = 1$. The generative model in xxx thus simplifies to

$$p\left(x_{1:T}, y_{1:T}, a, \lambda_x, b, \lambda_y\right) = p(a)p(\lambda_x)p(b)p\left(\lambda_y\right)\prod_{t=1}^{T} p(x_t|x_{t-1}, a, \lambda_x)p(y_t|x_t, b, \lambda_y) \tag{9.4}$$

The probability density functions on the right hand side of (9.4) are assumed to be given as

$$p(a) \coloneqq N(a; \mu_a, \sigma_a^2) = (2\pi\,\sigma_a^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_a^2}(a - \mu_a)^2\right) \tag{9.5}$$

$$p(\lambda_x) \coloneqq G\left(\lambda_x; a_{\lambda_x}, b_{\lambda_x}\right) = \frac{1}{\Gamma(a_{\lambda_x})}\frac{1}{b_{\lambda_x}{}^{a_{\lambda_x}}}\lambda_x{}^{a_{\lambda_x}-1} \exp\left(-\frac{\lambda_x}{b_{\lambda_x}}\right) \tag{9.6}$$

$$p(b) \coloneqq N\left(a; \mu_b, \sigma_b^2\right) = \left(2\pi\,\sigma_b^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_b^2}(b - \mu_b)^2\right) \tag{9.7}$$

$$p\left(\lambda_y\right) \coloneqq G\left(\lambda_y; a_{\lambda_y}, b_{\lambda_y}\right) = \frac{1}{\Gamma\left(a_{\lambda_y}\right)}\frac{1}{b_{\lambda_y}{}^{a_{\lambda_y}}}\lambda_y{}^{a_{\lambda_y}-1} \exp\left(-\frac{\lambda_y}{b_{\lambda_y}}\right) \tag{9.8}$$

$$p(x_t|x_{t-1}, a, \lambda_x) \coloneqq N(x_t|ax_{t-1}, \lambda_x^{-1}) = \left(\frac{\lambda_x}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_x}{2}(x_t - ax_{t-1})^2\right) \tag{9.9}$$

$$p\left(y_t|x_t, b, \lambda_y\right) \coloneqq N\left(x_t|bx_t, \lambda_y^{-1}\right) = \left(\frac{\lambda_y}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_y}{2}(y_t - bx_t)^2\right) \tag{9.10}$$

Next, we assume the following factorized form of the variational approximation to the posterior distribution over the unobserved variables

$$p\left(x_{1:T}, a, \lambda_x, b, \lambda_y|y_{1:T}\right) \approx q(x_{1:T})q(a)q(\lambda_x)q(b)q\left(\lambda_y\right) \tag{9.11}$$

were we set

$$q(x_{1:T}) \coloneqq \prod_{t=2}^{T} q(x_{t-1}, x_t) = \prod_{t=2}^{T} p(x_{t-1}, x_t|y_{1:T})\prod_{t=2}^{T} p\left(x_{t-1}, x_t; \mu_{x_{t-1:t}|y_{1:T}}, \Sigma_{x_{t-1:t}|y_{1:T}}\right) \tag{9.12}$$

and for VB-EM iterations $i = 0,1,2, \ldots$

$$q^{(i)}(a) \coloneqq N\left(a; \mu_a^{(i)}, \sigma_a^{2(i)}\right) = \left(2\pi\sigma_a^{2(i)}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_a^{2(i)}}\left(a - \mu_a^{(i)}\right)^2\right) \tag{9.13}$$

$$q^{(i)}(\lambda_x) \coloneqq G\left(\lambda_x; a_{\lambda_x}^{(i)}, b_{\lambda_x}^{(i)}\right) = \frac{1}{\Gamma\left(a_{\lambda_x}^{(i)}\right)} \frac{1}{b_{\lambda_x}^{(i)\,a_{\lambda_x}^{(i)}}} \lambda_x{}^{a_{\lambda_x}^{(i)}-1} \exp\left(-\frac{\lambda_x}{b_{\lambda_x}^{(i)}}\right) \tag{9.14}$$

$$q(b) \coloneqq N\left(a; \mu_b^{(i)}, \sigma_b^{2\,(i)}\right) = \left(2\pi\sigma_b^{2\,(i)}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_b^{2\,(i)}}\left(b - \mu_b^{(i)}\right)^2\right) \tag{9.15}$$

$$q\left(\lambda_y\right) \coloneqq G\left(\lambda_y; a_{\lambda_y}^{(i)}, b_{\lambda_y}^{(i)}\right) = \frac{1}{\Gamma\left(a_{\lambda_y}^{(i)}\right)} \frac{1}{b_{\lambda_y}^{(i)\,a_{\lambda_y}^{(i)}}} \lambda_y{}^{a_{\lambda_y}^{(i)}-1} \exp\left(-\frac{\lambda_y}{b_{\lambda_y}^{(i)}}\right) \tag{9.16}$$

Note that parameters of the prior and variational distributions are distinguished by means of the iteration superscript (i).

According to the variational Bayes theorem for factorized posteriors, the variational free energy is maximized for

$$q(\vartheta_i) \propto \exp\left(\int q\left(\vartheta_{\setminus i}\right) \ln p(y, \vartheta)\, d\vartheta_{\setminus i}\right) \tag{9.17}$$

For the current example, we have

$$y = y_{1:T} \text{ and } \vartheta \coloneqq \left\{x_{1:T}, a, \lambda_x, b, \lambda_y\right\} \tag{9.18}$$

where the unobserved variables partition according to

$$q(\vartheta) = \prod q(\vartheta_i) \coloneqq q(x_{1:T}) q(a) q(\lambda_x) q(b) q\left(\lambda_y\right) \tag{9.19}$$

We thus obtain the following update equations, using the short hand notation $\langle f(x)\rangle_{p(x)} \coloneqq \int p(x) f(x) dx$ for expectations

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\left\langle \ln p\left(y, x_{1:T}, a, \lambda_x, b, \lambda_y\right)\right\rangle_{q^{(i)}(a) q^{(i)}(\lambda_x) q^{(i)}(b) q^{(i)}(\lambda_y)}\right) \tag{9.20}$$

$$q^{(i+1)}(a) \propto \exp\left(\left\langle \ln p\left(y, x_{1:T}, a, \lambda_x, b, \lambda_y\right)\right\rangle_{q^{(i+1)}(x_{1:T}) q^{(i)}(\lambda_x) q^{(i)}(b) q^{(i)}(\lambda_y)}\right) \tag{9.21}$$

$$q^{(i+1)}(\lambda_x) \propto \exp\left(\left\langle \ln p\left(y, x_{1:T}, a, \lambda_x, b, \lambda_y\right)\right\rangle_{q^{(i)}(x_{1:T}) q^{(i)}(a) q^{(i)}(b) q^{(i)}(\lambda_y)}\right) \tag{9.22}$$

$$q^{(i+1)}(b) \propto \exp\left(\left\langle \ln p\left(y, x_{1:T}, a, \lambda_x, b, \lambda_y\right)\right\rangle_{q^{(i)}(x_{1:T}) q^{(i)}(a) q^{(i)}(\lambda_x) q^{(i)}(\lambda_y)}\right) \tag{9.23}$$

$$q^{(i+1)}\left(\lambda_y\right) \propto \exp\left(\left\langle \ln p\left(y, x_{1:T}, a, \lambda_x, b, \lambda_y\right)\right\rangle_{q^{(i)}(x_{1:T}) q^{(i)}(a) q^{(i)}(\lambda_x) q^{(i)}(b)}\right) \tag{9.24}$$

9.2 Evaluation of $\boldsymbol{q^{(i+1)}(x_{1:T})}$

Applying the parameter augmentations discussed in Supplement Section 8 to the univariate LGSSM, one obtains

$$\tilde{y}_t = \begin{pmatrix} y_t \\ 0 \\ 0 \end{pmatrix} \tag{9.25}$$

$$\tilde{a} = \langle a\rangle_{q(a)} = \mu_a$$

$$\tilde{\Sigma}_x = \left(\langle \lambda_x\rangle_{q(\lambda_x)}^{-1}\right)^{-1} = a_{\lambda_x} b_{\lambda_x}$$

$$\tilde{b} = \begin{pmatrix} \langle b \rangle_{q(b)} \\ u_A \\ u_B \end{pmatrix} = \begin{pmatrix} \langle b \rangle_{q(b)} \\ \sqrt{\langle a^2 \lambda_x \rangle_{q(a)q(\lambda_x)} - \langle a \rangle^2_{q(a)} \langle \lambda_x \rangle_{q(\lambda_x)}} \\ \sqrt{\langle b^2 \lambda_y \rangle_{q(b)q(\lambda_y)} - \langle b \rangle^2_{q(b)} \langle \lambda_y \rangle_{q(\lambda_y)}} \end{pmatrix} = \begin{pmatrix} \mu_b \\ \sqrt{(\mu_a + \sigma_a^2) a_{\lambda_x} b_{\lambda_x} - \mu_a^2 a_{\lambda_x} b_{\lambda_x}} \\ \sqrt{(\mu_b + \sigma_b^2) a_{\lambda_y} b_{\lambda_y} - \mu_b^2 a_{\lambda_y} b_{\lambda_y}} \end{pmatrix}$$

$$\tilde{\Sigma}_y = \begin{pmatrix} \left(\langle \lambda_y \rangle^{-1}_{q(\lambda_y)} \right)^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a_{\lambda_y} b_{\lambda_y} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

9.3 Evaluation of $q(a)q(\lambda_x)q(b)q(\lambda_y)$

To derive the update equations for the LGSSM, we take the following approach: We first reformulate the expectation of the joint distribution of the LGSSM and parameters, $p(y, x_{1:T}, a, \lambda_x, b, \lambda_y)$, under the variational mean field approximation $q^{(i+1)}(x_{1:T})q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)$. We can then in turn derive for each variational distribution by ignoring its contribution to the expectation based on the variational Bayes theorem.

○ *Step 1 – Evaluation of the variational expectation of* $p(y, x_{1:T}, a, \lambda_x, b, \lambda_y)$

$$\langle \ln p(y, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(x_{1:T})q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)} \tag{9.26}$$

$$= \langle \ln p(y, x_{1:T} | a, \lambda_x, b, \lambda_y) p(a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(x_{1:T})q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)}$$

Under the assumption of a factorized prior distribution

$$p(a, \lambda_x, b, \lambda_y) = p(a)p(\lambda_x)p(b)p(\lambda_y) \tag{9.27}$$

and with the likelihood

$$p(y, x_{1:T} | a, \lambda_x, b, \lambda_y) = \prod_{t=2}^{T} p(x_t | x_{t-1}, a, \lambda_x) \prod_{t=1}^{T} p(y_t | x_t, b, \lambda_y) \tag{9.28}$$

we obtain

$$\langle \ln p(y, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(x_{1:T})q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)} \tag{9.29}$$

$$= \sum_{t=2}^{T} \langle \ln p(x_t | x_{t-1}, a, \lambda_x) \rangle_{q^{(i+1)}(x_{t-1:t})q^{(i)}(a)q^{(i)}(\lambda_x)}$$

$$+ \sum_{t=1}^{T} \langle \ln p(y_t | x_t, b, \lambda_y) \rangle_{q^{(i+1)}(x_t)q^{(i)}(b)q^{(i)}(\lambda_y)}$$

$$+ \langle \ln p(a) \rangle_{q^{(i)}(a)}$$

$$+ \langle \ln p(\lambda_x) \rangle_{q^{(i)}(\lambda_x)}$$

$$+ \langle \ln p(b) \rangle_{q^{(i)}(b)}$$

$$+ \langle \ln p(\lambda_y) \rangle_{q^{(i)}(\lambda_y)}$$

Next, substitution the respective forms for the probability density functions, we obtain

$$\langle \ln p(y, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(x_{1:T})q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)} \tag{9.30}$$

$$= \sum_{t=2}^{T} \langle \ln \left(\left(\frac{\lambda_x}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda_x}{2} (x_t - a x_{t-1})^2 \right) \right) \rangle_{q^{(i+1)}(x_{t-1:t}) q^{(i)}(a) q^{(i)}(\lambda_x)}$$

$$+ \sum_{t=1}^{T} \langle \ln \left(\left(\frac{\lambda_y}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda_y}{2} (y_t - b x_t)^2 \right) \right) \rangle_{q^{(i+1)}(x_t) q^{(i)}(b) q^{(i)}(\lambda_y)}$$

$$+ \langle \ln \left(\left(\frac{1}{2\pi\sigma_a^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\sigma_a^2} \left(a - \mu_a^{(i)} \right)^2 \right) \right) \rangle_{q^{(i)}(a)}$$

$$+ \langle \ln \left(\frac{1}{\Gamma\left(a_{\lambda_x}^{(i)} \right)} \frac{1}{\left(b_{\lambda_x}^{(i)} \right)^{a_{\lambda_x}^{(i)}}} \lambda_x^{\, a_{\lambda_x}^{(i)} - 1} \exp \left(-\frac{\lambda_x}{b_{\lambda_x}^{(i)}} \right) \right) \rangle_{q^{(i)}(\lambda_x)}$$

$$+ \langle \ln \left(\left(\frac{1}{2\pi\sigma_b^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\sigma_b^2} \left(b - \mu_b^{(i)} \right)^2 \right) \right) \rangle_{q^{(i)}(b)}$$

$$+ \langle \ln \left(\frac{1}{\Gamma\left(a_{\lambda_y}^{(i)} \right)} \frac{1}{\left(b_{\lambda_y}^{(i)} \right)^{a_{\lambda_y}^{(i)}}} \lambda_y^{\, a_{\lambda_y}^{(i)} - 1} \exp \left(-\frac{\lambda_y}{b_{\lambda_y}^{(i)}} \right) \right) \rangle_{q^{(i)}(\lambda_y)}$$

Evaluation of the logarithmic expressions then yields

$$\langle \ln p\left(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y \right) \rangle_{q^{(i+1)}(x_{1:T}) q^{(i)}(a) q^{(i)}(\lambda_x) q^{(i)}(b) q^{(i)}(\lambda_y)} \tag{9.31}$$

$$= \frac{T-1}{2} \langle \ln \left(\frac{\lambda_x}{2\pi} \right) \rangle_{q^{(i)}(\lambda_x)} - \frac{1}{2} \sum_{t=2}^{T} \langle \lambda_x (x_t - a x_{t-1})^2 \rangle_{q^{(i+1)}(x_{t-1:t}) q^{(i)}(a) q^{(i)}(\lambda_x)}$$

$$+ \frac{T}{2} \langle \ln \left(\frac{\lambda_y}{2\pi} \right) \rangle_{q^{(i)}(\lambda_y)} - \frac{1}{2} \sum_{t=1}^{T} \langle \lambda_y (y_t - b x_t)^2 \rangle_{q^{(i+1)}(x_{t-1:t}) q^{(i)}(b) q^{(i)}(\lambda_y)}$$

$$- \langle \frac{1}{2\sigma_a^2} \left(a - \mu_a^{(i)} \right)^2 \rangle_{q^{(i)}(a)} - \frac{1}{2} \ln 2\pi\sigma_a^2$$

$$- \ln \left(\Gamma \left(a_{\lambda_x}^{(i)} \right) \right) - \ln \left(b_{\lambda_x}^{(i)} \right)^{a_{\lambda_x}^{(i)}} + \left(a_{\lambda_x}^{(i)} - 1 \right) \langle \ln \lambda_x \rangle_{q(\lambda_x)} - \frac{1}{b_{\lambda_x}^{(i)}} \langle \lambda_x \rangle_{q(\lambda_x)}$$

$$- \langle \frac{1}{2\sigma_b^2} \left(b - \mu_b^{(i)} \right)^2 \rangle_{q^{(i)}(b)} - \frac{1}{2} \ln 2\pi\sigma_b^2$$

$$- \ln \left(\Gamma \left(a_{\lambda_y}^{(i)} \right) \right) - \ln \left(b_{\lambda_y}^{(i)} \right)^{a_{\lambda_x}^{(i)}} + \left(a_{\lambda_y}^{(i)} - 1 \right) \langle \ln \lambda_y \rangle_{q(\lambda_y)} - \frac{1}{b_{\lambda_y}^{(i)}} \langle \lambda_y \rangle_{q(\lambda_y)}$$

Evaluation of the quadratic terms then yields

$$\langle \ln p\left(y, x_{1:T}, a, \lambda_x, b, \lambda_y \right) \rangle_{q^{(i+1)}(x_{1:T}) q^{(i)}(a) q^{(i)}(\lambda_x) q^{(i)}(b) q^{(i)}(\lambda_y)} \tag{9.32}$$

$$= \frac{T-1}{2} \langle \ln \left(\frac{\lambda_x}{2\pi} \right) \rangle_{q^{(i)}(\lambda_x)} - \frac{1}{2} \langle \lambda_x \left(\sum_{t=2}^{T} x_t x_t - 2a \sum_{t=2}^{T} x_{t-1} x_t + a^2 \sum_{t=2}^{T} x_{t-1} x_{t-1} \right) \rangle_{q^{(i+1)}(x_{t-1:t}) q^{(i)}(a) q^{(i)}(\lambda_x)}$$

$$+ \frac{T}{2} \langle \ln \left(\frac{\lambda_y}{2\pi} \right) \rangle_{q^{(i)}(\lambda_y)} - \frac{1}{2} \langle \lambda_y \left(\sum_{t=1}^{T} y_t y_t - 2b \sum_{t=2}^{T} y_t x_t + b^2 \sum_{t=1}^{T} x_t x_t \right) \rangle_{q^{(i+1)}(x_t) q^{(i)}(b) q^{(i)}(\lambda_y)}$$

$$- \frac{1}{2\sigma_a^2} \langle a^2 - 2a\mu_a^{(i)} + \left(\mu_a^{(i)} \right)^2 \rangle_{q^{(i)}(a)} - \frac{1}{2} \ln 2\pi\sigma_a^2$$

$$-\ln\left(\Gamma\left(a_{\lambda_x}^{(i)}\right)\right)-\ln\left(b_{\lambda_x}^{(i)}\right)^{a_{\lambda_x}^{(i)}}+\left(a_{\lambda_x}^{(i)}-1\right)\langle\ln\lambda_x\rangle_{q(\lambda_x)}-\frac{1}{b_{\lambda_x}^{(i)}}\langle\lambda_x\rangle_{q(\lambda_x)}$$

$$-\frac{1}{2\sigma_b^2}\langle b^2-2b\mu_b^{(i)}+\left(\mu_b^{(i)}\right)^2\rangle_{q^{(i)}(b)}-\frac{1}{2}\ln 2\pi\sigma_b^2$$

$$-\ln\left(\Gamma\left(a_{\lambda_y}^{(i)}\right)\right)-\ln\left(b_{\lambda_y}^{(i)}\right)^{a_{\lambda_x}^{(i)}}+\left(a_{\lambda_y}^{(i)}-1\right)\langle\ln\lambda_y\rangle_{q(\lambda_y)}-\frac{1}{b_{\lambda_y}^{(i)}}\langle\lambda_y\rangle_{q(\lambda_y)}$$

- *Step 2 – Application of the Variational Bayes Inference Theorem*

Based on the above, we now apply

$$p(\vartheta_i)\propto\exp\left(\int q\left(\vartheta_{\backslash i}\right)\ln p(y,\vartheta)\,d\vartheta_{\backslash i}\right) \tag{9.33}$$

for each variational parameter distribution $q(a), q(\lambda_x), q(b)$, and $q(\lambda_y)$. Likewise, we ignore all terms in (9.33) which are independent of the variable under consideration, because these terms influence the respective variational probabilities independent of the variable under consideration.

Update equation for $q(a)$

We have

$$\ln q(a)\propto-\frac{1}{2}\langle\sum_{t=2}^{T}\lambda_x x_t x_t-2a\sum_{t=2}^{T}\lambda_x x_{t-1}x_t+a^2\sum_{t=2}^{T}\lambda_x x_{t-1}x_{t-1}\rangle_{q^{(i+1)}(x_{t-1:t})q^{(i)}(\lambda_x)}-\frac{1}{2\sigma_a^2}\left(a^2-2a\mu_a^{(i)}+\left(\mu_a^{(i)}\right)^2\right) \tag{9.34}$$

Partitioning the expectations, we obtain

$$\ln q(a)\propto-\frac{1}{2}\sum_{t=2}^{T}\langle\lambda_x x_t x_t\rangle_{q(x_t)q(\lambda_x)}-2a\sum_{t=2}^{T}\langle\lambda_x x_{t-1}x_t\rangle_{q(x_{t-1:t})q(\lambda_x)}+a^2\sum_{t=2}^{T}\langle\lambda_x x_{t-1}x_{t-1}\rangle_{q(x_{t-1})q(\lambda_x)}-\frac{1}{2\sigma_a^2}\left(a^2-2a\mu_a^{(i)}+\left(\mu_a^{(i)}\right)^2\right) \tag{9.35}$$

Rewriting the above as a quadratic form in a yields and ignoring terms independent of a yields

$$\ln q(a)\propto-\frac{1}{2}\left(\sum_{t=2}^{T}\langle\lambda_x x_{t-1}x_{t-1}\rangle_{q(x_{t-1})q(\lambda_x)}+\frac{1}{2\sigma_a^2}\right)a^2-\left(\sum_{t=2}^{T}\langle\lambda_x x_{t-1}x_t\rangle_{q(x_{t-1:t})q(\lambda_x)}+\frac{\mu_a^{(i)}}{\sigma_a^2}\right)a \tag{9.36}$$

According to the completing-the-square theorem (Supplement Section 2),

$$\exp\left(-\frac{1}{2}\alpha a^2-\beta a\right)\propto N(a;\alpha^{-1}\beta,\alpha^{-1}) \tag{9.37}$$

we thus have

$$q^{(i+1)}(a)\propto N\left(a;\mu_a^{(i+1)},\sigma_a^{2\,(i+1)}\right) \tag{9.38}$$

where

$$\sigma_a^{2\,(i+1)}\coloneqq\left(\sum_{t=2}^{T}\langle\lambda_x x_{t-1}x_{t-1}\rangle_{q^{(i+1)}(x_{t-1})q(\lambda_x)}+\frac{1}{\sigma_a^{2\,(i)}}\right)^{-1} \tag{9.39}$$

and

$$\mu_a^{(i+1)} \coloneqq \sigma_a^{2\,(i+1)} \left(\sum_{t=2}^{T} \langle \lambda_x x_{t-1} x_t \rangle_{q(x_{t-1:t}) q(\lambda_x)} + \frac{\mu_a^{(i)}}{\sigma_a^2} \right) \tag{9.40}$$

Update equation for $q(\lambda_x)$

From (9.33), we have, using the expectation partitioning as above and reordering in terms of $\ln \lambda_x$ and λ_x in we have

$$\ln q^{(i+1)}(\lambda_x) \propto \left(\frac{T-1}{2} + a_{\lambda_x}^{(i)} - 1 \right) \ln \lambda_x - \left(\frac{1}{2} \left(\sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)} - 2 \sum_{t=2}^{T} \langle a x_{t-1} x_t \rangle_{q(x_{t-1:t}) q(a)} + \sum_{t=2}^{T} \langle a^2 x_{t-1} x_{t-1} \rangle_{q(a) q(x_{t-1})} \right) + \frac{1}{b_{\lambda_x}^{(i)}} \right) \lambda_x \tag{9.41}$$

Taking the exponential on both sides then yields

$$q^{(i+1)}(\lambda_x) \propto \lambda_x^{\left(\frac{T-1}{2} + a_{\lambda_x}^{(i)} - 1 \right)} \exp\left(- \left(\frac{1}{b_{\lambda_x}^{(i)}} + \frac{1}{2} \left(\sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)} - 2 \sum_{t=2}^{T} \langle a x_{t-1} x_t \rangle_{q(x_{t-1:t}) q(a)} + \sum_{t=2}^{T} \langle a^2 x_{t-1} x_{t-1} \rangle_{q(a) q(x_{t-1})} \right) \right) \lambda_x \right) \tag{9.42}$$

Up to a normalization constant, $q^{(i+1)}(\lambda_x)$ is thus given by a Gamma distribution

$$q^{(i+1)}(\lambda_x) \propto \lambda^{a_{\lambda_x}^{(i+1)}} \exp\left(- \frac{\lambda}{b_{\lambda_x}^{(i)}} \right) \tag{9.43}$$

with

$$a_{\lambda_x}^{(i+1)} = \frac{T-1}{2} + a_{\lambda_x}^{(i)} \tag{9.44}$$

and

$$b_{\lambda_x}^{(i)} = \left(\frac{1}{b_{\lambda_x}^{(i)}} + \frac{1}{2} \left(\sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)} - 2 \sum_{t=2}^{T} \langle a x_{t-1} x_t \rangle_{q(x_{t-1:t}) q(a)} + \sum_{t=2}^{T} \langle a^2 x_{t-1} x_{t-1} \rangle_{q(a) q(x_{t-1})} \right) \right)^{-1} \tag{9.45}$$

Update equation for $q(b)$

We have

$$\ln q(b) \propto -\frac{1}{2} \lambda_y \langle \sum_{t=1}^{T} y_t y_t - 2b \sum_{t=1}^{T} y_t x_t + b^2 \sum_{t=1}^{T} x_t x_t \rangle_{q^{(i+1)}(x_{t-1:t}) q^{(i)}(\lambda_y)} - \frac{1}{2\sigma_b^2} \left(b^2 - 2b\mu_b^{(i)} + \left(\mu_b^{(i)} \right)^2 \right) \tag{9.46}$$

Partitioning the expectations, we obtain

$$\ln q(b) \propto -\frac{1}{2} \lambda_y \sum_{t=1}^{T} y_t y_t - 2b\lambda_y \sum_{t=1}^{T} y_t \langle x_t \rangle_{q(x_t)} + b^2 \lambda_y \sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)} - \frac{1}{2\sigma_b^2} \left(b^2 - 2b\mu_b^{(i)} + \left(\mu_b^{(i)} \right)^2 \right) \tag{9.47}$$

Rewriting the above as a quadratic form in b yields and ignoring terms independent of b yields

$$\ln q(b) \propto -\frac{1}{2} \left(\lambda_y \sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)} + \frac{1}{2\sigma_b^2} \right) b^2 - \left(\lambda_y \sum_{t=1}^{T} y_t \langle x_t \rangle_{q(x_t)} + \frac{\mu_b^{(i)}}{\sigma_b^2} \right) b \tag{9.48}$$

According to the completing-the-square theorem,

$$\exp\left(-\frac{1}{2}\alpha b^2 - \beta b\right) \propto N(b; \alpha^{-1}\beta, \alpha^{-1}) \tag{9.49}$$

we thus have

$$q^{(i+1)}(b) \propto N\left(b; \mu_a^{(i+1)}, \sigma_b^{2\,(i+1)}\right) \tag{9.50}$$

where

$$\sigma_b^{2\,(i+1)} \coloneqq \left(\lambda_y \sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)} + \frac{1}{2\sigma_b^2}\right)^{-1} \tag{9.51}$$

and

$$\mu_b^{(i+1)} \coloneqq \sigma_b^{2\,(i+1)} \left(\lambda_y \sum_{t=1}^{T} y_t \langle x_t \rangle_{q(x_t)} + \frac{\mu_b^{(i)}}{\sigma_b^2}\right) \tag{9.52}$$

Update equation for $q(\lambda_y)$

From (9.33), we have, using the expectation partitioning as above and reordering in terms of $\ln \lambda_x$ and λ_x in we have

$$\ln q^{(i+1)}(\lambda_y) \propto \left(\frac{T-1}{2} + a_{\lambda_y}^{(i)} - 1\right) \ln \lambda_y - \left(\frac{1}{2}\left(\sum_{t=1}^{T} y_t y_t - 2b\lambda_y \sum_{t=1}^{T} y_t \langle x_t \rangle_{q(x_t)} + b^2 \lambda_y \sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)}\right) + \frac{1}{b_{\lambda_y}^{(i)}}\right) \lambda_y \tag{9.53}$$

Taking the exponential on both sides then yields

$$q^{(i+1)}(\lambda_y) \propto \lambda_y^{\left(\frac{T-1}{2} + a_{\lambda_y}^{(i)} - 1\right)} \exp\left(-\left(\frac{1}{b_{\lambda_y}^{(i)}} + \frac{1}{2}\left(\sum_{t=1}^{T} y_t y_t - 2b\lambda_y \sum_{t=1}^{T} y_t \langle x_t \rangle_{q(x_t)}\right) + b^2 \lambda_y \sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)}\right) \lambda_y\right) \tag{9.54}$$

Up to a normalization constant, $q^{(i+1)}(\lambda_x)$ is thus given by a Gamma distribution

$$q^{(i+1)}(\lambda_y) \propto \lambda^{a_{\lambda_y}^{(i+1)}} \exp\left(-\frac{\lambda}{b_{\lambda_y}^{(i)}}\right) \tag{9.55}$$

with

$$a_{\lambda_y}^{(i+1)} = \frac{T-1}{2} + a_{\lambda_y}^{(i)} \tag{9.56}$$

and

$$b_{\lambda_y}^{(i)} = \left(\frac{1}{b_{\lambda_y}^{(i)}} + \frac{1}{2}\left(\sum_{t=1}^{T} y_t y_t - 2b\lambda_y \sum_{t=1}^{T} y_t \langle x_t \rangle_{q(x_t)}\right) + b^2 \lambda_y \sum_{t=2}^{T} \langle x_t x_t \rangle_{q(x_t)}\right)^{-1} \tag{9.57}$$

9.4 Evaluation of $\mathcal{F}(q(\vartheta))$

As in Supplement Section 6, we have

$$\mathcal{F}(q(\vartheta)) = \int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta + \mathcal{H}(q(\vartheta)) - \mathcal{KL}(q(\vartheta)||p(\vartheta)) \tag{9.58}$$

Evaluation of the average energy term $\int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta$

In the current context, the average energy term is given as

$$\int q(\vartheta) \ln p(y|\vartheta)\, d\vartheta = \langle \ln p\big(y_{1:T}|x_{1:T}, a, \lambda_x, b, \lambda_y\big)\rangle_{q(x_{1:T})q(a)q(\lambda_x)q(b)q(\lambda_y)} \tag{9.59}$$

$$= \langle \ln\big(\prod_{t=1}^{T} p\big(y_t\big|x_t, b, \lambda_y\big) \prod_{t=2}^{T} p(x_t|x_{t-1}, a, \lambda_x)\big)\rangle_{q(x_{1:T})q(a)q(\lambda_x)q(b)q(\lambda_y)}$$

$$= \langle \sum_{t=1}^{T} \ln p\big(y_t\big|x_t, b, \lambda_y\big)\rangle_{q(x_{1:T})q(b)q(\lambda_y)} + \langle \sum_{t=1}^{T} \ln p(x_t|x_{t-1}, a, \lambda_x)\rangle_{q(x_{1:T})q(a)q(\lambda_x)}$$

$$= \sum_{t=1}^{T} \langle \ln\left(\left(\frac{\lambda_y}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_y}{2}(y_t - bx_t)^2\right)\right)\rangle_{q(x_{1:T})q(b)q(\lambda_y)}$$

$$+ \sum_{t=1}^{T} \langle \ln\left(\left(\frac{\lambda_x}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_x}{2}(x_t - ax_{t-1})^2\right)\right)\rangle_{q(x_{1:T})q(a)q(\lambda_x)}$$

$$= \sum_{t=1}^{T} \langle \frac{1}{2}\ln\lambda_x - \frac{1}{2}\ln 2\pi - \frac{\lambda_x}{2}x_t^2 - ax_t x_{t-1} - a^2 x_t^2\rangle_{q(x_{1:T})q(a)q(\lambda_x)}$$

$$+ \sum_{t=1}^{T} \langle \frac{1}{2}\ln\lambda_y - \frac{1}{2}\ln 2\pi - \frac{\lambda_y}{2}y_t^2 - y_t bx_t - b^2 x_t^2\rangle_{q(x_{1:T})q(b)q(\lambda_y)}$$

$$= \frac{T}{2}\langle \ln\lambda_x\rangle_{q(\lambda_x)} - \frac{1}{2}\sum_{t=1}^{T}\langle \lambda_x x_t^2\rangle_{q(x_t)q(\lambda_x)}$$

$$- \sum_{t=1}^{T}\langle ax_t x_{t-1}\rangle_{q(a)q(x_{t-1:t})} - \sum_{t=1}^{T}\langle a^2 x_t^2\rangle_{q(a)q(x_t)} - \frac{T}{2}\ln 2\pi$$

$$+ \frac{T}{2}\langle \ln\lambda_y\rangle_{q(\lambda_y)} - \frac{1}{2}\sum_{t=1}^{T} y_t^2\langle \lambda_y\rangle_{q(\lambda_y)}$$

$$- \sum_{t=1}^{T} y_t\langle bx_t\rangle_{q(b)q(x_t)} - \sum_{t=1}^{T}\langle b^2 x_t^2\rangle_{q(b)q(x_t)} - \frac{T}{2}\ln 2\pi$$

Considering the remaining integral terms in turn then yields

a)

$$\langle \ln\lambda_x\rangle_{q(\lambda_x)} = \int q(\lambda_x)\ln\lambda_x\, dx = \big(\psi\big(a_{\lambda_x}\big) + \ln b_{\lambda_x}\big) \tag{9.60}$$

b)

$$\langle \lambda_x x_t^2\rangle_{q(\lambda_x)q(x_t)} = \iint q(\lambda_x)q(x_t)\lambda_x x_t^2\, dx_t d\lambda_x \tag{9.61}$$

$$= \left(\int\left(\int q(\lambda_x)q(x_t)\lambda_x x_t^2\, dx_t\right) d\lambda_x\right)$$

$$= \left(\int\left(q(\lambda_x)\lambda_x \int q(x_t)x_t^2\, dx_t\right) d\lambda_x\right)$$

$$= \left(\int\left(q(\lambda_x)\lambda_x(\mu_{x_t|y_{1:T}}^2 + \sigma_{x_t|y_{1:T}}^2)\right) d\lambda_x\right)$$

$$= (\mu_{x_t|y_{1:T}}^2 + \sigma_{x_t|y_{1:T}}^2)\int q(\lambda_x)\lambda_x\, d\lambda_x$$

$$= a_{\lambda_x} b_{\lambda_x} (\mu^2_{x_t|y_{1:T}} + \sigma^2_{x_t|y_{1:T}})$$

c)

$$\langle a x_t x_{t-1} \rangle_{q(a)q(x_{t-1:t})} = \iint q(a) q(x_{t-1:t}) a x_t x_{t-1} da dx_{t-1:t} \tag{9.62}$$

$$= \int (\int q(a) q(x_{t-1:t}) a x_t x_{t-1} \, da) \, dx_{t-1:t}$$

$$= \int (q(x_{t-1:t}) x_t x_{t-1} \int q(a) a \, da) \, dx_{t-1:t}$$

$$= \mu_a (\int q(x_{t-1:t}) x_t x_{t-1} \, dx_{t-1:t})$$

$$= \mu_a (\int (\int q(x_{t-1:t}) x_t x_{t-1} \, dx_{t-1}) \, dx_t)$$

$$= \mu_a (\int (q(x_t) x_t \int q(x_{t-t}) x_{t-1} \, dx_{t-1}) \, dx_t)$$

$$= \mu_a \mu_{x_t|y_{1:T}} \mu_{x_{t-1}|y_{1:T}}$$

d)

$$\langle a^2 x_t^2 \rangle_{q(a)q(x_t)} = \iint q(a) q(x_t) a^2 x_t^2 da dx_t \tag{9.63}$$

$$= \int (\int q(a) q(x_t) a^2 x_t^2 da) dx_t$$

$$= \int (q(x_t) x_t^2 \int q(a) a^2 da) dx_t$$

$$= (\mu_a^2 + \sigma_a^2) \int (q(x_t) x_t^2) dx_t$$

$$= (\mu_a^2 + \sigma_a^2)(\mu^2_{x_t|y_{1:T}} + \sigma^2_{x_t|y_{1:T}})$$

e)

$$\langle \ln \lambda_y \rangle_{q(\lambda_y)} = \left(\psi \left(a_{\lambda_y} \right) + \ln b_{\lambda_y} \right) \tag{9.64}$$

f)

$$\langle \lambda_y \rangle_{q(\lambda_y)} = a_{\lambda_y} b_{\lambda_y} \tag{9.65}$$

g)

$$\langle b x_t \rangle_{q(x_t)q(b)} = \iint q(x_t) q(b) b x_t \, db dx_t \tag{9.66}$$

$$= \int (\int q(x_t) q(b) b x_t \, db) dx_t$$

$$= \int (q(x_t) x_t \int q(b) b \, db) dx_t$$

$$= \mu_b \int (q(x_t) x_t) dx_t$$

$$= \mu_b \mu_{x_t|y_{1:T}}$$

h)

$$\langle b^2 x_t^2 \rangle_{q(x_t)q(b)} = \iint q(x_t)q(b)b^2 x_t^2 \, dbdx_t \tag{9.67}$$

$$= \int (\int q(x_t)q(b)b^2 x_t^2 \, db)dx_t$$

$$= \int (q(x_t)x_t^2 \int q(b)b^2 \, db)dx_t$$

$$= (\mu_b^2 + \sigma_b^2) \int (q(x_t)x_t^2)dx_t$$

$$= (\mu_b^2 + \sigma_b^2)(\mu_{x_t|y_{1:T}}^2 + \sigma_{x_t|y_{1:T}}^2)$$

Substitution of results a)-h) into (9.58) then yields an expression of the average energy in terms of the parameters governing the unobserved variables and the data as follows:

$$\langle \ln p(y_{1:T}|x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q(x_{1:T})q(a)q(\lambda_x)q(b)q(\lambda_y)} = \frac{T}{2}\left(\left(\psi(a_{\lambda_x}) + \ln b_{\lambda_x}\right)\right) \tag{9.68}$$

$$-\frac{1}{2}\sum_{t=1}^{T}(a_{\lambda_x} b_{\lambda_x}(\mu_{x_{t|y_{1:T}}}^2 + \sigma_{x_t|y_{1:T}}^2))$$

$$-\sum_{t=1}^{T}(\mu_a \mu_{x_t|y_{1:T}} \mu_{x_{t-1}|y_{1:T}})$$

$$-\sum_{t=1}^{T}((\mu_a^2 + \sigma_a^2)(\mu_{x_t|y_{1:T}}^2 + \sigma_{x_t|y_{1:T}}^2))$$

$$-\frac{T}{2}\ln 2\pi$$

$$+\frac{T}{2}\left(\psi\left(a_{\lambda_y}\right) + \ln b_{\lambda_y}\right)$$

$$-\frac{1}{2}\sum_{t=1}^{T} y_t^2 a_{\lambda_y} b_{\lambda_y}$$

$$-\sum_{t=1}^{T} y_t \mu_b \mu_{x_t|y_{1:T}}$$

$$-\sum_{t=1}^{T}(\mu_b^2 + \sigma_b^2)(\mu_{x_t|y_{1:T}}^2 + \sigma_{x_t|y_{1:T}}^2)$$

$$-\frac{T}{2}\ln 2\pi$$

$$= \frac{T}{2}\left(\left(\psi(a_{\lambda_x}) + \ln b_{\lambda_x}\right) + \psi\left(a_{\lambda_y}\right) + \ln b_{\lambda_y}\right)$$

$$-\frac{a_{\lambda_x} b_{\lambda_x}}{2}\sum_{t=1}^{T} \mu_{x_{t|y_{1:T}}}^2 + \sigma_{x_t|y_{1:T}}^2$$

$$-\sum_{t=1}^{T}(\mu_a \mu_{x_t|y_{1:T}} \mu_{x_{t-1}|y_{1:T}})$$

$$-(\mu_a^2+\sigma_a^2)\sum_{t=1}^{T}\mu_{x_t|y_{1:T}}^2+\sigma_{x_t|y_{1:T}}^2$$

$$-\frac{a_{\lambda_y}b_{\lambda_y}}{2}\sum_{t=1}^{T}y_t^2$$

$$-\mu_b\sum_{t=1}^{T}y_t\mu_{x_t|y_{1:T}}$$

$$-(\mu_b^2+\sigma_b^2)\sum_{t=1}^{T}(\mu_{x_t|y_{1:T}}^2+\sigma_{x_t|y_{1:T}}^2)$$

$$-T\ln 2\pi$$

Evaluation of the entropy term $\mathcal{H}(q(\vartheta))$

Due to the factorization properties of the variational approximation, the entropy term $\mathcal{H}\big(\mathrm{q}(\vartheta)\big)$ decomposes into a sum of entropies as follows:

$$\mathcal{H}\big(q(\vartheta)\big)=\mathcal{H}\left(q(x_{1:T})q(a)q(\lambda_x)q(b)q\big(\lambda_y\big)\right) \tag{9.69}$$

$$=\mathcal{H}\big(q(x_{1:T})\big)+\mathcal{H}\big(q(a)\big)+\mathcal{H}(\lambda_x)+\mathcal{H}\big(q(b)\big)+\mathcal{H}(q(\lambda_y))$$

Evaluating the terms in (9.69) in turn then yields

$$\mathcal{H}\big(q(x_{1:T})\big)=\mathcal{H}(\textstyle\prod_{t=2}^{T}q(x_{t-1},x_t)) \tag{9.70}$$

$$=\textstyle\sum_{t=2}^{T}\mathcal{H}\big(q(x_{t-1},x_t)\big)$$

$$=\textstyle\sum_{t=2}^{T}\mathcal{H}\big(N\big(x_{t-1:t};\mu_{x_{t-1:t}|y_{1:T}},\Sigma_{x_{t-1:t}|y_{1:T}}\big)\big)$$

The entropy of the unobserved variables $x_{1:T}$ is thus given as the sum of the differential entropies of adjacent variables x_{t-1} and x_t, i.e. as the sum of the differential entropies of $T-1$ bivariate Gaussian distributions. We thus have

$$\mathcal{H}\big(q(x_{1:T})\big)=(T-1)(1+\ln 2\pi)+\tfrac{1}{2}\textstyle\sum_{t=2}^{T}\ln\big|\Sigma_{x_{t-1:t}|y_{1:T}}\big| \tag{9.71}$$

Further, we have

$$\mathcal{H}\big(q(a)\big)=\mathcal{H}\big(N(a;\mu_a,\sigma_a^2)\big)=\tfrac{1}{2}\ln\sigma_a^2+\tfrac{1}{2}(1+\ln(2\pi)) \tag{9.72}$$

$$\mathcal{H}(q(\lambda_x))=\mathcal{H}\big(G(\lambda_x;a_{\lambda_x},b_{\lambda_x}\big)=\ln\Gamma(a_{\lambda_x})-\big(a_{\lambda_x}-1\big)\psi\big(a_{\lambda_x}\big)-\ln b_{\lambda_x}+a_{\lambda_x} \tag{9.73}$$

$$\mathcal{H}\big(q(b)\big)=\mathcal{H}\left(N\big(b;\mu_b,\sigma_b^2\big)\right)=\tfrac{1}{2}\ln\sigma_b^2+\tfrac{1}{2}(1+\ln(2\pi)) \tag{9.74}$$

$$\mathcal{H}\left(q\big(\lambda_y\big)\right)=\mathcal{H}\left(G(\lambda_y;a_{\lambda_y},b_{\lambda_y}\right)=\ln\Gamma(a_{\lambda_y})-\left(a_{\lambda_y}-1\right)\psi\left(a_{\lambda_y}\right)-\ln b_{\lambda_y}+a_{\lambda_y} \tag{9.75}$$

Evaluation of the KL divergence term $\mathcal{KL}\big(q(\vartheta)||p(\vartheta)\big)$

We first note that due to the factorization properties of $q(\vartheta)$, the KL divergence term can be rewritten as the sum of KL divergence terms over unobserved variable partitions as follows

$$\mathcal{KL}\big(q(\vartheta)||p(\vartheta)\big) = \mathcal{KL}\left(q(x_{1:T})q(a)q(\lambda_x)q(b)q\big(\lambda_y\big)||p(x_{1:T})p(a)p(\lambda_x)p(b)p\big(\lambda_y\big)\right) \tag{9.76}$$

$$= \mathcal{KL}\big(q(x_{1:T})||p(x_{1:T})\big)$$

$$+\mathcal{KL}(q(a)||p(a)) + \mathcal{KL}(q(\lambda_x)||p(\lambda_x))$$

$$+\mathcal{KL}(q(b)||p(b)) + \mathcal{KL}\big(q\big(\lambda_y\big)||p(\lambda_y)\big)$$

Using the results in (Penny, 2001), we can then evaluate the respective KL divergence terms in turn as follows

$$\mathcal{KL}\big(q(x_{1:T})||p(x_{1:T})\big) = \mathcal{KL}(\prod_{t=2}^{T} q(x_{t-1}, x_t) \,||\, \prod_{t=2}^{T} p(x_{t-1}, x_t)) \tag{9.77}$$

$$= \sum_{t=2}^{T} \mathcal{KL}\big(q(x_{t-1}, x_t)||p(x_{t-1}, x_t)\big)$$

$$= \sum_{t=2}^{T} \mathcal{KL}\left(N\left(x_{t-1:t}; \mu^{(i)}_{x_{t-1:t}|y_{1:T}}, \Sigma^{(i)}_{x_{t-1:t}|y_{1:T}}\right)||N\left(x_{t-1:t}; \mu^{(0)}_{x_{t-1:t}|y_{1:T}}, \Sigma^{(0)}_{x_{t-1:t}|y_{1:T}}\right)\right)$$

Further, we have

$$\mathcal{KL}(q(a)||p(a)) = \mathcal{KL}\left(N\left(a; \mu_a^{(i)}, \sigma_a^{2\,(i)}\right)||N\left(a; \mu_a^{(0)}, \sigma_a^{2\,(0)}\right)\right) \tag{9.78}$$

$$\mathcal{KL}(q(\lambda_x)||p(\lambda_x)) = \mathcal{KL}\left(G\left(\lambda_x; a_{\lambda_x}^{(i)}, b_{\lambda_x}^{(i)}\right)||G\left(\lambda_x; a_{\lambda_x}^{(0)}, b_{\lambda_x}^{(0)}\right)\right) \tag{9.79}$$

$$\mathcal{KL}(q(b)||p(b)) = \mathcal{KL}\left(N\left(a; \mu_b^{(i)}, \sigma_b^{2\,(i)}\right)||N\left(a; \mu_b^{(0)}, \sigma_b^{2\,(0)}\right)\right) \tag{9.80}$$

$$\mathcal{KL}\big(q\big(\lambda_y\big)||p(\lambda_y)\big) = \mathcal{KL}\left(G\left(\lambda_y; a_{\lambda_y}^{(i)}, b_{\lambda_y}^{(i)}\right)||G\left(\lambda_y; a_{\lambda_y}^{(0)}, b_y^{(0)}\right)\right) \tag{9.81}$$

9.5 Probability density function transformations

To estimate a posterior distribution over the latent SDE parameters $\alpha \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$, we used the Euler-Maruyama approximation, which entails the use of two transformation mappings from α and σ onto the LGSSM parameter $a \in \mathbb{R}$ and $\sigma_x^2 \in \mathbb{R}_+$, as discussed in Supplement Section 3. If we obtain posterior pdfs over a and σ_x^2 based on VB inference for the LGSSM and would like to use them for posterior inference over the SDE parameters α and σ, we have to respect the transformation theorem for pdfs, as discussed below.

General transformation theorem

In general, we have for a probability measure P on the Borel σ-field $\mathcal{B}$ with pdf f of x and a measurable linear-affine transformation given by

$$T: \mathbb{R} \to \mathbb{R}, y \coloneqq T(x) \quad (x \in \mathbb{R}^n) \tag{9.82}$$

the following pdf g for y

$$g(y) = \frac{f\left(T^{-1}(y)\right)}{\left|T'\left(T^{-1}(y)\right)\right|} \tag{9.83}$$

Transformation theorem for pdf over a to pdf over α

For the special univariate linear affine transformation case, we have with $a, b \in \mathbb{R}, a \neq 0$

$$T: \mathbb{R} \to \mathbb{R}, y \coloneqq T(x) \coloneqq ax + b \ (x \in \mathbb{R}) \tag{9.84}$$

the following pdf g for y:

$$g(y) = \frac{1}{|a|} \cdot f\left(\frac{y-b}{a}\right) \ \ (y \in \mathbb{R}) \tag{9.85}$$

Above, we have seen that the inverse mapping of the transformation from a to α is given by

$$T_1: \mathbb{R} \to \mathbb{R}, a \mapsto T_1(a) = \alpha = \frac{1}{\Delta t}(a-1) = \frac{1}{\Delta t}a - \frac{1}{\Delta t} \tag{9.86}$$

Substitution thus yields for the posterior pdf g over α, given the posterior pdf f over a

$$g(\alpha) = \Delta t \cdot f\big(\Delta t(\alpha + \Delta t)\big) \ \ (\alpha \in \mathbb{R}) \tag{9.87}$$

Transformation theorem for pdf over λ_x to pdf over σ

Above, we have seen that

$$T_2: \mathbb{R}_+ \to \mathbb{R}_+, \lambda_x \mapsto T_2(\lambda_x) = \Delta t \lambda_x^{-\frac{1}{2}} \tag{9.88}$$

with derivative

$$T_2'(\lambda_x) = -\frac{1}{2}\Delta t \lambda_x^{-\frac{3}{2}} \tag{9.89}$$

and inverse

$$T_2^{-1}: \mathbb{R}_+ \to \mathbb{R}_+, \sigma \mapsto T_2^{-1}(\sigma) \coloneqq \left(\frac{\Delta t}{\sigma}\right)^2 \tag{9.90}$$

Substitution in (9.83) thus yields

$$g(\sigma) = \frac{f\left(\left(\frac{\Delta t}{\sigma}\right)^2\right)}{\left|-\frac{1}{2}\Delta t\left(\left(\frac{\Delta t}{\sigma}\right)^2\right)^{-\frac{3}{2}}\right|} = \frac{2 \cdot f\left(\left(\frac{\Delta t}{\sigma}\right)^2\right)}{\Delta t\left(\frac{\Delta t}{\sigma}\right)^{-3}} = \frac{2 \cdot f\left(\left(\frac{\Delta t}{\sigma}\right)^2\right)}{\Delta t \frac{\sigma^3}{\Delta t^3}} = \frac{2\Delta t^2}{\sigma^3} \cdot f\left(\left(\frac{\Delta t}{\sigma}\right)^2\right) \tag{9.91}$$

Transformation theorem for pds over λ_y to pdf over σ_y^2

Obviously, we have

$$T_3: \mathbb{R}_+ \to \mathbb{R}_+, \lambda_x \mapsto T_3(\lambda_x) = \lambda_x^{-1} \tag{9.92}$$

with derivative

$$T_3'(\lambda_x) = -\lambda_x^{-2} \tag{9.93}$$

and inverse

$$T_3^{-1}: \mathbb{R}_+ \to \mathbb{R}_+, \sigma_y^2 \mapsto T_3^{-1}\left(\sigma_y^2\right) = \left(\sigma_y^2\right)^{-1} \tag{9.94}$$

Substitution in (9.83) thus yields

$$g\left(\sigma_y^2\right) = \frac{f\left(\left(\sigma_y^2\right)^{-1}\right)}{\left|-\left(\left(\sigma_y^2\right)^{-1}\right)^{-2}\right|} = \left(\sigma_y^2\right)^{-2} f\left(\left(\sigma_y^2\right)^{-1}\right) \tag{9.95}$$

References

Barber, D., & Chiappa, S. (2007). Unified inference for variational Bayesian linear Gaussian state-space model. *In: Schölkopf, B. and Platt, P. and Hofmann, T., (eds.) Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. (pp. 81-88).* The MIT Press: Cambridge, US.

Barber, David. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Bass, R. F. (2011). *Stochastic Processes*. Cambridge University Press.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning* (1st ed. 2006. Corr. 2nd printing.). Springer New York.

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Briers, M., Doucet, A., & Maskell, S. (2010). Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, *62*(1), 61–89. doi:10.1007/s10463-009-0236-2

Bucy, R. S., & Joseph, P. D. (1987). *Filtering for Stochastic Processes With Applications to Guidance*. Chelsea Publishing Company.

Cover, T. M., & Thomas. (1991). *Elements of information theory*. New York: Wiley.

Hassler, U. (2007). *Stochastische Integration Und Zeitreihenmodellierung: Eine Einfuhrung Mit Anwendungen Aus Finanzierung Und Okonometrie* Springer, New York.

Johari, H., Desabrais, K. J., Rauch, H. E., Striebel, C. T., & Tung, F. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, *3*(8), 1445–1450. doi:10.2514/3.3166

Kalman, R. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering*, (82 (Series D)), 35–45.

Kloeden, P. E., & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations* (1st ed. 1992. Corr. 4th printing.). Springer, New York.

Michael Chappell, Adrian Groves and Mark Woolrich. (2008). *TR08MC1 : The FMRIB Variational Bayes Tutorial*. Oxford. Available from http://users.fmrib.ox.ac.uk/~chappell/papers/TR07MC1.pdf

Penny, W. D. (2001). *Kullback-Liebler Divergences of Normal, Gamma, Dirichlet and Wishart Densities. Technical report, Wellcome Department of Cognitive Neurology, 2001*. Available from *www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps*

Richter, Mathias. (o. J.). Variationsrechnung. *Variationsrechnung*. Available from http://me-lrt.de/kategorie/master/variationsrechnung

W. D. Penny, S. J. R. (2000). Variational Bayes for 1-dimensional Mixture Models. Available from www.robots.ox.ac.uk/~sjrob/Pubs/vbmog.ps.gz

Yu, B., Shenoy, K., & Sahani, M. (2004). Derivation of Kalman Filtering and Smoothing Equations. Available from http://www-npl.stanford.edu/~byronyu/papers/derive_ks.pdf